

# **SAS® Content Categorization Studio 5.2 Quick Start Guide**



The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2011.  
*SAS® Content Categorization Studio 5.2: Quick Start Guide*. Cary, NC: SAS Institute Inc.

### **SAS® Content Categorization Studio 5.2: Quick Start Guide**

Copyright © 2011, SAS Institute Inc., Cary, NC, USA

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

**For a hard-copy 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 Restricted Rights Notice:** 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.

1st electronic book, July 2011

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](http://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

---

<b>About This Book .....</b>	<b>1</b>
Audience .....	1
Prerequisites .....	1
Conventions .....	1
 <b>1 About SAS Content Categorization Studio .....</b>	 <b>3</b>
1.1 What Is SAS Content Categorization Studio? .....	3
1.2 Benefits of Using SAS Content Categorization Studio .....	4
1.3 How Does SAS Content Categorization Studio Work? .....	5
1.4 About the Architecture .....	5
 <b>2 Creating a Categorizer .....</b>	 <b>7</b>
2.1 Overview of Creating a Categorizer .....	7
2.2 Creating a New Project .....	8
2.2.1 Overview of Project Creation .....	8
2.2.2 Create a New Project .....	9
2.2.3 Specify Installation-Specific Operations .....	11
2.2.4 Specify Project Settings .....	14
2.2.4.A Overview of Project Settings .....	14
2.2.4.B Specify Project Settings for Categories .....	15
2.2.4.C Specify the Miscellaneous Project Settings .....	18
2.3 Add Categories and Their Metadata .....	20
2.4 Troubleshooting Project Development .....	22
 <b>3 Writing Category Rules .....</b>	 <b>23</b>
3.1 Overview of Writing Category Rules .....	23
3.2 Writing Linguistic Rules .....	24
3.2.1 Write a List of Unique Identifiers .....	24
3.2.2 Selecting Special Symbols .....	25
3.2.3 Weighted Linguistic Rules .....	27
3.3 Writing Boolean Rules .....	28
3.3.1 Overview of Boolean Rules .....	28
3.3.2 Adding Boolean Operators .....	28
3.3.3 Adding Special Symbols .....	31
3.3.4 Expanding Word Forms .....	32

---

3.3.5 Specifying a Structured Text Field .....	33
3.3.6 Viewing Boolean Rules .....	33
3.4 Build and Save the Categorizer .....	35
<b>4 Testing and Uploading Category Rules .....</b>	<b>37</b>
4.1 Overview of Testing and Uploading Category Rules .....	37
4.2 Testing Categories .....	38
4.2.1 Create a Testing Directory .....	38
4.2.2 Populate Testing Directories .....	40
4.2.3 Test the Category Rule .....	40
4.3 Use the Graphical Reports .....	41
4.4 Test a Document in the Document Window .....	45
4.5 Test against All Testing Documents .....	46
4.6 Test Failing Test Files .....	48
4.7 Upload the Categorizer to SAS Content Categorization Server .....	48
<b>Appendixes .....</b>	<b>51</b>
<b>Recommended Reading .....</b>	<b>53</b>
<b>Glossary .....</b>	<b>55</b>
<b>Index .....</b>	<b>57</b>

---

# About This Book

---

## Audience

This book is designed for advanced users of SAS Content Categorization Studio who want a quick overview of the operations that are necessary to create a project. This book assumes the underlying technological expertise that is necessary to quickly develop a taxonomy of categories with few detailed instructions.

## Prerequisites

Here are the prerequisites for using SAS Content Categorization Studio:


- Load SAS Content Categorization Studio onto your machine.
- Obtain access to documents that are representative of the types of texts that you plan to categorize.
- If you plan to input Web pages, make sure that a supported browser is loaded onto your machine.
- If you create a project that uses a UTF-8 language, install the prerequisite fonts.

## Conventions

This manual uses the following typographical conventions:

Convention	Description
<code>(OR, _tmac: "@Composers")</code>	The code examples for Boolean rules are shown in a fixed-width font.
<b>Browse</b> button	The labels for user interface controls are shown in a bold, sans-serif font.
Top	The names of taxonomy nodes appear in a fixed-width font.

---

Convention	Description
<a href="http://www.sas.com">www.sas.com</a>	The hypertext links are shown in a light blue, fixed-width font, and are underlined.
	The Question Mark button accesses <i>SAS Content Categorization Studio: User's Guide</i> in PDF format.

---

# Chapter: 1

## About SAS Content Categorization Studio

---

- *What Is SAS Content Categorization Studio?*
- *Benefits of Using SAS Content Categorization Studio*
- *How Does SAS Content Categorization Studio Work?*
- *About the Architecture*

### 1.1 What Is SAS Content Categorization Studio?

In most organizations it is necessary to obtain information about, and from, data that is created internally and externally. SAS Content Categorization Studio enables you to define a taxonomy of categories that identify matching documents.

This Quick Start Guide is limited to a discussion of how to create a project using categories. You can also use this application to develop a concepts branch of the taxonomy. For more information about concepts, see *SAS Content Categorization Studio: User's Guide*.

Using an intuitive, Windows interface, users with various skill sets and levels of expertise can develop a taxonomy. You can then write rules for the categories that classify data.

Easy taxonomy creation

Use the **Taxonomy** tab to create a visual taxonomy. This taxonomy has branches for different languages if you are building one project that uses multiple languages.

---

---

### Easy rule development

Use the **Rules** tab to write a category rule and click the **Syntax Check** button in this window to validate the syntax of the rule. *Rule* is used within this guide to refer to the syntax that defines category membership for input documents.

### Easy Testing

Test your rules using groups of 10-20 documents that you assemble into a testing taxonomy, or choose to create a central repository of files. You can also collect documents that should fail and place them into a separate fail directory for testing purposes. For example, the word *bush* in landscaping documents should not match the *President Bush* category.

### Easy Uploading

After you develop and test the taxonomy, you can upload the compiled taxonomy rules as a .mco binary file to SAS Content Categorization Server. The category rules in this file are automatically applied to incoming documents.

## 1.2 Benefits of Using SAS Content Categorization Studio

SAS Content Categorization Studio provides users with the following benefits:

Empower subject matter experts and taxonomists by providing a simple, visual user interface where you build a taxonomy, define rules, and test:

SAS Content Categorization Studio includes easy-to-use windows that simplify large, complex, and hierarchical taxonomies. You can specify your own rules, test, and generate .mco files. These files are applied by SAS Content Categorization Server to input documents.

Develop metadata for your information:

SAS Content Categorization Studio uses advanced linguistic technologies to identify metadata in, and about, your documents.



---

Improve the business value of information technology and the corporate data that it manages:

SAS Content Categorization Studio creates .mco files that automate the classification and extraction of entities from input documents during real time using SAS Content Categorization Server.

Save money on information retrieval and organization costs:

All of the information created by, or within, your organization can be classified and located. You can find information that is related, whether you know the exact terms that you are seeking.

## 1.3 How Does SAS Content Categorization Studio Work?

SAS Content Categorization Studio is a Windows application that anyone can use to develop taxonomies that classify and extract the information found in your organization. Interactively identify the data that you need without using a programming language.

You can upload the output .mco file to SAS Content Categorization Server where this file is automatically applied to input documents.

## 1.4 About the Architecture

Use the figure below to understand the processes used during the following two phases:

Management phase

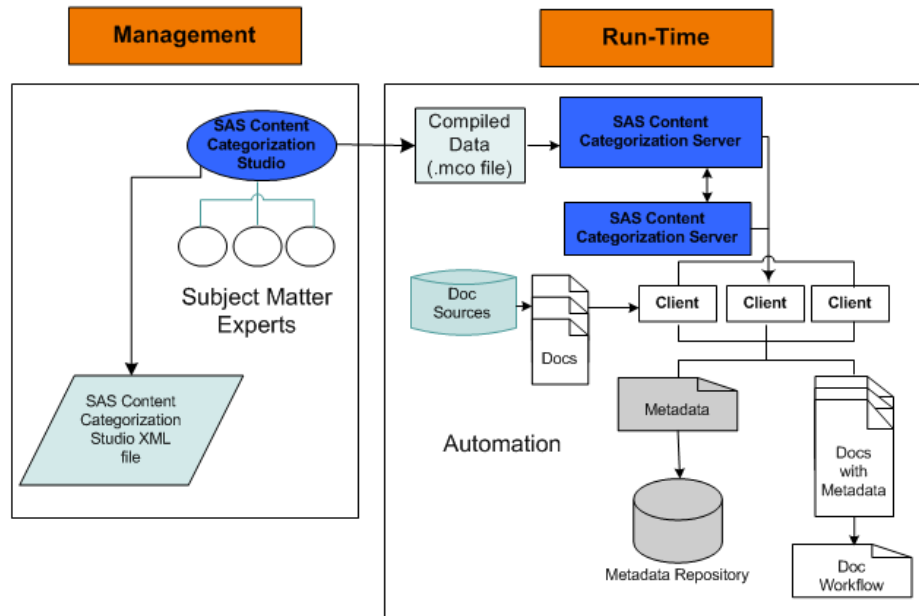
Subject matter experts specify a taxonomy of categories. During the second part of this phase, these experts write rules to ensure that all of the documents that should match a category are located. This is known as *recall*. These rules should also ensure *precision*, meaning that those texts that should not match are not returned as matches for the selected taxonomy node.

---

## Run time

The compiled SAS Content Categorization Studio data (.mco file) is sent to the SAS Content Categorization Server. SAS Content Categorization Server returns data about the input documents such as what categories the document matched.

*Figure 1-1 Architecture*



---

# 2

## Creating a Categorizer

---

- *Overview of Creating a Categorizer*
- *Creating a New Project*
- *Add Categories and Their Metadata*
- *Troubleshooting Project Development*

### 2.1 Overview of Creating a Categorizer

In most organizations, the information that is produced needs to be organized, or classified, according to its subject matter. SAS Content Categorization Studio enables you to customize this classification structure. You can return data about your information, or metadata, by using the SAS Content Categorization Studio solution.

Categorization is based on rules that determine the matching category for input documents. Together these categories form a taxonomy that organizes the subject matter areas. The categorizer classifies input documents into this taxonomy. For example, travel articles about former President Bush's summer home in Kennebunkport, ME could be categorized under the *Travel* category. In this case, texts describing how to plant bushes should not match the *Travel* category.

The SAS Content Categorization Studio testing process uses the testing taxonomy to determine the precision and recall of your categorizer. *Precision* measures the relevancy of the matched documents, while *recall* measures whether all of the texts that should be returned are matched. For these reasons, each category rule should be broad enough to include all of the texts that you expect to match. These rules should also exclude any documents that do not belong to the selected category.

To gain an overview of categorizer development, read through these steps:

- 
1. Specify the categories that comprise the taxonomy, or the overall classification scheme, of your project. The categories are determined by the subject matter in the corpus of documents that you plan to categorize. For more information, see Section 2.2 *Creating a New Project* on page 8.
  2. Write the rules that match the subject matter of input documents. Accurate rules match all of the relevant documents, but do not match irrelevant texts. For example, all of the input documents covering the *Rangers hockey players* should match the *Sports* category, while those on *park rangers* should match the *Parks* category. For more information, see Chapter 3.
  3. Build your categorizer as you develop the taxonomy and write your rules. For more information, see Section 3.4 *Build and Save the Categorizer* on page 35.
  4. Test the rules against a set of testing files to ensure the adequacy of precision and recall. In other words, by selecting small groups of documents that you expect to pass and placing them into a testing directory that matches your taxonomy, you can test your taxonomy. For more information, see Chapter 4.
  5. Upload the categorizer to SAS Content Categorization Server.
  6. (Optional) Query an index using the Boolean rules that you write for your categories.

## 2.2 Creating a New Project

### 2.2.1 Overview of Project Creation

This section assumes that you are creating a project of linguistic and Boolean category rules. For this reason, the steps necessary to develop a statistical or the automatic rule generator tool for categories, or for concept extraction, is not included in this book.

---

## 2.2.2 Create a New Project

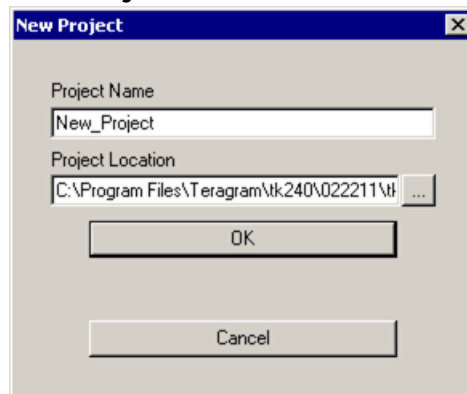
This section assumes that you have installed SAS Content Categorization Studio and opened the user interface.


To create a new project, complete these steps:

1. Select **Start --> Programs --> SAS Content Categorization Studio** and the untitled user interface appears.

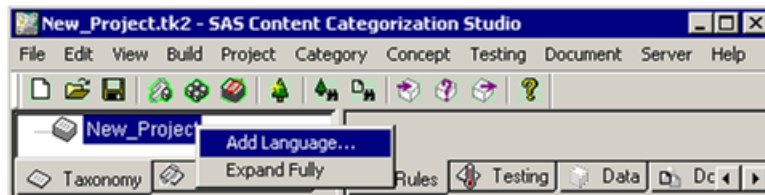



2. Select **File --> New Project**.

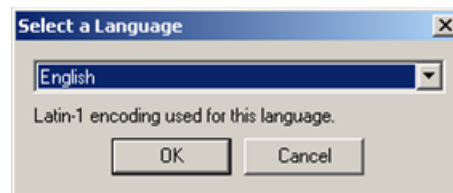


3. Enter the following information into the New Project window:
  - a. Enter the name of the new project into the **Project Name** field.
  - b. By default, the path to the projects folder is automatically entered into the **Project Location** field. Click  to locate a different folder for this project.
  - c. Click **OK**.

- 
4. Right-click on the project name node that appears in the Taxonomy window. For example, select `New_Project`. Select **Add Language** from the drop-down menu that appears.

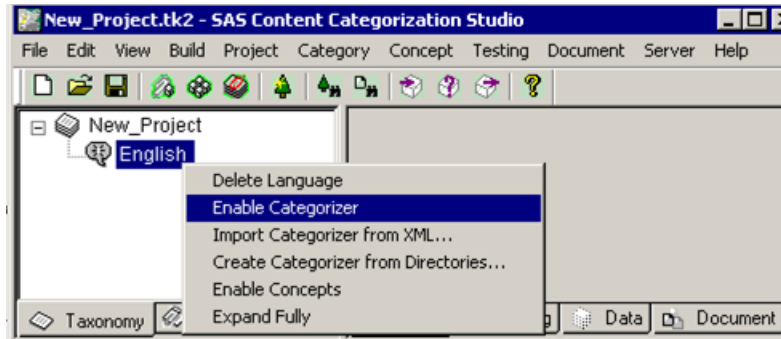


5. The Select a Language window appears.
  - a. Click  to select a language for your taxonomy. If you purchased more than one language, all of these languages appear in the drop-down list.
  - b. Make sure that you selected the correct language encoding. See the message that appears in the Select a Language window below the selected language. For example, see **Latin-1 encoding used for this language**.



- c. Click **OK**.
6. (Optional) Continue using the Select a Language window to add languages that you purchased. Each language creates a separate branch in the taxonomy.

- 
7. Right-click on the language icon that appears and select **Enable Categorizer** from the drop-down menu that appears.

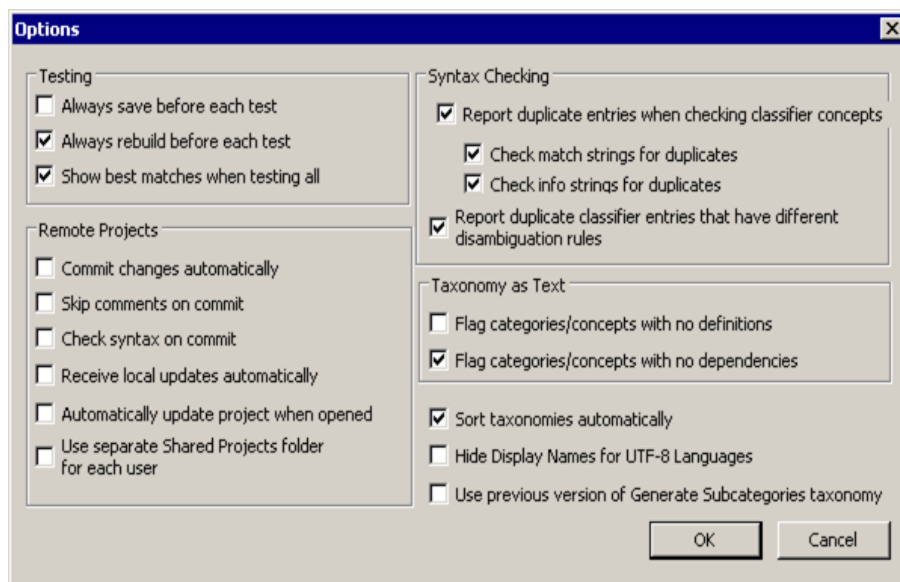


### 2.2.3 Specify Installation-Specific Operations

The Options window enables you to specify operations that apply to all of the projects in the installation, until you reset these selections.

To use the Options window to specify these operations, complete these steps:

1. Select **Edit --> Options** to set the installation-specific operations for SAS Content Categorization Studio. The default settings are selected below:



2. (Optional) Under **Testing** choose any of the following operations:

**Always save before each test**

Save the project before each testing operation.

**Always rebuild before each test**

(Default) Deselect if you do not want to automatically build the project binary files before each test.

**Show best matches when testing all**

(Default) The Best Matches window appears when you test **All categories** in the **Document** tab. Deselect if you do not require this data.

---

**Hint:** The operations that are available under **Remote Projects** are available only after you also install SAS Content Categorization Collaborative Server.

---



- 
3. (Optional) Under **Syntax Checking** choose from the following operations:

**Report duplicate entries when checking classifier concepts**

Choose one, or both of the following selections:

**Check match strings for duplicates**

Examine the match part of the classifier concept definition.

**Check info strings for duplicates**

Examine the information part of the classifier concept definition.

**Report duplicate classifier entries that have different disambiguation rules**

Locate duplicates. Also select either, or both, of the selections above.

4. (Optional) Under **Taxonomy as Text**, select either or both of the following operations:

**Flag categories/concepts with no definitions**

See the categories and concepts without rules and definitions in the Notepad window that appears.

**Flag categories/concepts with no dependencies**

(Default) See the categories and concepts that share rules in the Notepad window that appears.

5. (Default) If you deselect **Sort taxonomies automatically**, each branch of the taxonomy is not alphabetically sorted.

---

**Note:** Click the plus sign (+) to the left of the *Categorizer*, *Concepts*, *Top*, *language*, and *project name* nodes. This action enables you to see the reordered taxonomy after each of these nodes is closed and reopened.

---

6. (Applies to UTF-8 languages, only) Select **Hide Display Names for UTF-8 Languages** to display the Latin-1 internal category names, while the UTF-8 names are hidden. This operation works in coordination with the Enter Names window.

- 
7. (Optional) Select **Use previous version of Generate Subcategories taxonomy** to use the previous version of the taxonomy instead of the Wikipedia taxonomy.
  8. Click **OK**.

## 2.2.4 Specify Project Settings

### 2.2.4.A Overview of Project Settings

Project settings, unlike the operations that are available in the Options window, apply only to the currently selected branch of the open project.

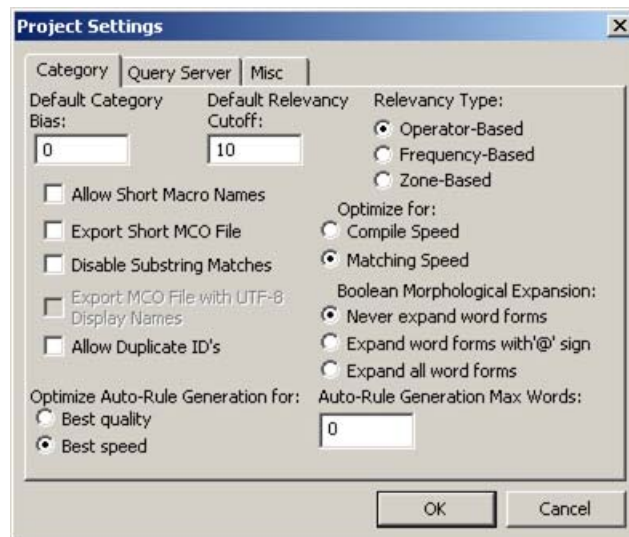
To open the Project Settings window, complete this step:

Select **Project --> Settings**. Use the Project Settings windows to set taxonomy-wide operations. If you choose to develop a SAS Content Categorization Studio project that uses more than one language, set the project settings for each language taxonomy separately.

---

**Notes:** Settings for automatic rule generation and statistical categorization are not covered in this guide.  
The **Query Server** Project Settings tab is not covered.

---



---

## 2.2.4.B Specify Project Settings for Categories

Use the **Category** tab of the Project Settings window to specify the operations that apply to the categories in the open project:

1. (Optional) Choose the following operations in the **Category** tab to customize the results returned by your project:

### Default Category Bias

(Default: 0) Assign more weight to your categories in the field. Use this setting to boost the relevancy of your categories into the range used by a third-party software application. If a number is entered into this field, this number is multiplied by the **Category Bias** setting in the **Data** tab.

### Default Relevancy Cutoff

(Default: 0) Specify the minimum relevancy required for a document to be a match on the selected category in the field. This setting applies to each of the relevancy types unless you specified another number for the **Relevancy Cutoff** field in the **Data** tab.

2. Specify the type of relevancy that is used to determine the category that is the best match for an input document under the **Relevancy Type** heading:

### Operator-Based

(Default) Deselect if you want to use another relevancy type. Leave this default selection to specify Boolean operators in your rules.

### Frequency-Based

Specify that the number of matching terms in a document determines the degree of its relevancy for a specific category.

### Zone-Based

Weight matches that occur in certain sections of an input document more heavily than matches in other areas.

3. (Boolean rules, only) Enable the use of short macro names in Boolean rules when you select **Allow Short Macro Names**. You specify macro names with the `_tmac` symbol such as:

(OR, `_tmac:"@Top/Music/Baroque/Composers"`)

By default, the unique name of a category is its full path such as:

---

Top/Music/Baroque/Composers

4. Choose the **Short Macro Names** operation so that you can refer to the short form of the category name in a macro rule. For example, you can specify the following syntax:

(OR, \_tmac: "@Composers")

5. Produce a \*.short.mco file when you select **Export Short MCO file**. This is a categorization binary file where the category names that are returned are the short paths, instead of the full pathnames.
6. Prevent a partial match on a string that defines a category rule when you select **Disable Substring Matches**. For example, if *business processes* and *business* are specified in the rule, a match is not returned for the word *business*. This is true unless the word *processes* immediately follows *business*.
7. Build a taxonomy using a UTF-8 language **Export MCO file with UTF-8 Display Names**. Use this operation to display the UTF-8 names in the category binary file. An additional <language>.mco file is created in the following format:

<language>.utf8.mco

The .mco file contains the Latin-1 internal names. The <language>.utf8.mco file enables you to see the taxonomy in the UTF-8 language that appears in the **Taxonomy** tab. For example, if you create a taxonomy structure of categories using Japanese, you might see the following line of text instead of Top/School:

Top/学校

8. Enable duplicate identification numbers to be entered into the ID field of the **Data** tab for categories when you select **Allow Duplicate ID's**. Otherwise, ensure that the identification numbers are unique.
9. (Boolean rules, only) Select one of these choices to specify the type of word form expansion under **Boolean Morphological Expansion**:

**Never expand word forms**

(Default) Matches occur only on the words that explicitly appear in rules. Words that are followed by an at sign (@) are treated as literals to be matched. For example, run@ matches only run@ in an

---

incoming text. If the words *run* and *running* also appear in this text stream, they are not matched.

**Expand word forms with '@' sign**

Expand only the words followed by an @ sign during the compile operation. The expanded forms appear in the .mco file. When the word ends with the following symbols, expansion is applied as described below:

@: both noun and verb forms

@v: verb forms only

@N: noun forms only

**Expand all word forms**

Treat every word in a category rule as if it ended with an @ sign when your project is compiled.

---

**Note:** You can use the **Expand Forms** button in the **Rules** tab to see any expansions that you are unsure about before you compile your project. In this case, select **Edit --> Undo** to return the @ signs.

---

10. Leave the default setting **Best Speed** under the **Optimize Auto-Rule for** heading unless you are building a taxonomy with thousands of nodes. In this case, select **Best Quality**.
11. (default: 0, means that there is no upper bound on the number of words returned as a match) Specify a maximum number of words or phrases for an automatically generated rule in the **Auto-Rule Generation Max Words** field.
12. Click **OK**.

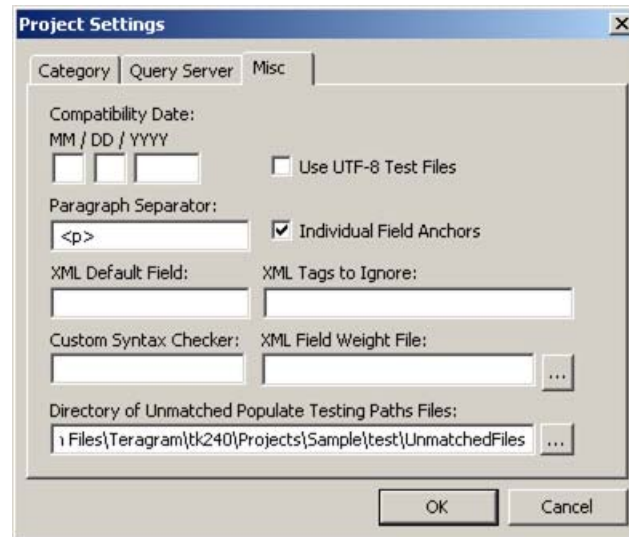
---

### 2.2.4.C Specify the Miscellaneous Project Settings

Use the **Misc** tab of the Project Settings window to specify the miscellaneous operations that apply to the open project:

To use the **Misc** tab, complete these steps:

1. Select the **Misc** tab.



2. Enter the date of the older version of SAS Content Categorization Server that you are running into the MM/DD/YYYY fields. This action sets the **Compatibility Date** for the .mco file that is automatically generated by SAS Content Categorization Studio. Use this field only with deprecated versions of this application. SAS Content Categorization Studio generates a binary file (.mco) that is compatible with the older version of SAS Content Categorization Server. Use this operation until you have time to install and run a newer version of SAS Content Categorization Server.
3. Select **Use UTF-8 Test Files** field when your testing documents are in UTF-8 format, but the language of the categorizer might not be UTF-8.

- 
4. (only for the rule-based categorizer that uses Boolean rules and for some concepts) Enter the string that is used as a paragraph separator within your documents into the **Paragraph Separator** field. For example, type <P>.
  5. Select **Individual Field Anchors** when you write Boolean rules that use disambiguation. By default, if you have more than one instance of the same XML tag in a Web document, SAS Content Categorization Studio collapses the sections into one searchable area. When you select this check box, each section of a Web-based document is searched separately. This feature has implications for some Boolean operators. For more information, see Section 11.8.4 *How to Use Project Settings with Structured Text* on page 318.
  6. Enter one or more XML fields into **XML Default Field** when you write your category rule. This operation limits search to the specified field instead of all of the XML fields in the input Web document.
  7. Enter one or more XML fields into **XML Tags to Ignore**. This operation excludes one or more XML fields when Web-based documents are processed.
  8. Specify the path to an external, custom grammar checker program in the **Custom Syntax Checker Executable** field that is used in place of the internal syntax checker program.
  9. Enter the path of a text file that weights the XML tags in input documents in the **XML Field Weight File** field.
  10. Enter the path to the testing documents that do not match any categories in your taxonomy into **Directory for Unmatched Populate Files**.

---

## 2.3 Add Categories and Their Metadata

After you create a new project, you can continue to define this project by adding categories to the taxonomy.

To add categories to the taxonomy, complete these steps:

1. Right-click on the **Top** node in the Taxonomy window and select **Add Category** from the drop-down menu that appears.
2. Type the name of the new category into the box that appears to the right of the new node.

The screenshot shows the 'New Project.tk2 - SAS Content Categorization Studio' window. On the left, a tree view shows the project structure: 'New\_Project' > 'English' > 'Categorizer' > 'Top' > 'New'. The main panel displays the 'New' category configuration. It includes fields for 'ID' (1344278), 'Author' (MyName), and 'Created' (February 23, 2011). There are also fields for 'Relevancy Cutoff' (10), 'Relevancy Bias' (10), 'Category Bias' (0), and 'Match Ratio' (0). Radio buttons are present for 'Completed' (selected), 'Pending', and 'Test Disabled'. Below these are text areas for 'Description' (The New category is used to explain the process of adding categories), 'Thesaurus', 'Query', 'Comments' (This category is provided only for sample purposes), and 'Related Links'. At the bottom, there are 'Testing Path' and 'Training Path' fields, each with a 'Propagate' button. To the right of these paths are 'Propagate Options' with checkboxes for 'Identical Path' and 'Create Folders'. The bottom status bar shows tabs for 'Taxonomy', 'Rules', 'Testing', 'Data', and 'Document'.

3. Specify the metadata for each category using the Data window.

---

**Hint:** All of these settings, or changes to the default specifications, are optional.

---



- 
- a. Enter the identification number that you use to track each category into the **ID** field.
  - b. Enter the name of the person who developed the category into the **Author** field.
  - c. The **Created** and **Modified** dates are automatically filled in for you.
  - d. Enter the number for the minimum threshold for frequency-based ranking into the **Relevancy Cutoff** field. Unless this number of instances of matching terms occurs in an input document, a match does not occur.
  - e. (Default: 1) Enter the number that is multiplied by all of the relevancy scores for this category into the **Relevancy Bias field**. Use this setting to boost the relevancy of this category in relation to the other categories in the taxonomy. This setting applies to both linguistic and Boolean rules and is used when third-party software is not a concern.
  - f. (Default: 0) Specify a number in the **Category Bias** field that is multiplied by the **Default Category Bias** setting in the Project Settings - Category window. This number is used only with third-party software and for rules that are defined by one term.
  - g. (Default: 10%) Specify the percentage of matching terms that make an input document a match for the category in the **Match Ratio**. This setting is used internally to convert linguistic rules to Boolean rules.
- 

**Note:** The +, \*\*, and -- symbols override the match ratio setting.

---

- h. (Default: **Completed**) Select a different status for the category rule. Choose **Pending** if you are still working on the category. Select **Test Disabled** if you are creating a helper category and do not want the rule to be matched.
- i. Enter a brief summary of the category into the **Description** field.

- 
- j. Enter a comma-separated (,) list of words that are alternative names (synonyms) for the category into the **Thesaurus** field. A search on an alternate name matches this category.
  - k. Enter a search term to locate matching documents in an index into the **Query** field.
  - l. Enter explanations or notes into the **Comments** field.
  - m. Enter a list of URLs that contain related information into the **Related Links** field.
  - n. Enter the pathname of the directory that contains the testing documents for this category into the **Testing Path** field.

Alternatively, click  to locate this directory.

- o. Use the **Propagate Options** to load testing documents into the testing directories. For more information, see Chapter 4.

## 2.4 Troubleshooting Project Development

To prevent unexpected results from occurring, consider the following points:

- To ensure accurate precision and recall, build part of the taxonomy and test before developing the entire taxonomy.
- Build the categorizer frequently as you develop your project and add rules. For more information, see Chapter 3.
- Save your project often.
- Test the categorizer reiteratively to check your settings and your rules. For more information, see Chapter 4.

---

# 3

## Writing Category Rules

---

- *Overview of Writing Category Rules*
- *Writing Linguistic Rules*
- *Writing Boolean Rules*
- *Build and Save the Categorizer*

### 3.1 Overview of Writing Category Rules

Write category rules to determine how categories are matched in input documents. There are two types of rules that you can write. However, linguistic rules are applied as Boolean rules in SAS Content Categorization Server

#### Linguistic rules

Write lists of unique terms that identify category members. SAS Content Categorization Studio analyzes each category rule in the context of the entire taxonomy before it matches an input document to a category. For this reason, if the list of identifying terms for a category is not unique, precision and recall are not optimized.

#### Boolean rules

(Recommended selection) Qualify unique, identifying terms with Boolean operators. Use this type of rule to limit the location of the matches. For example, specify the location where a matching term is located in an input document.

---

**Note:** Your taxonomy can contain categories using different types of rules, but you can specify only one type of rule for each category.

---

---

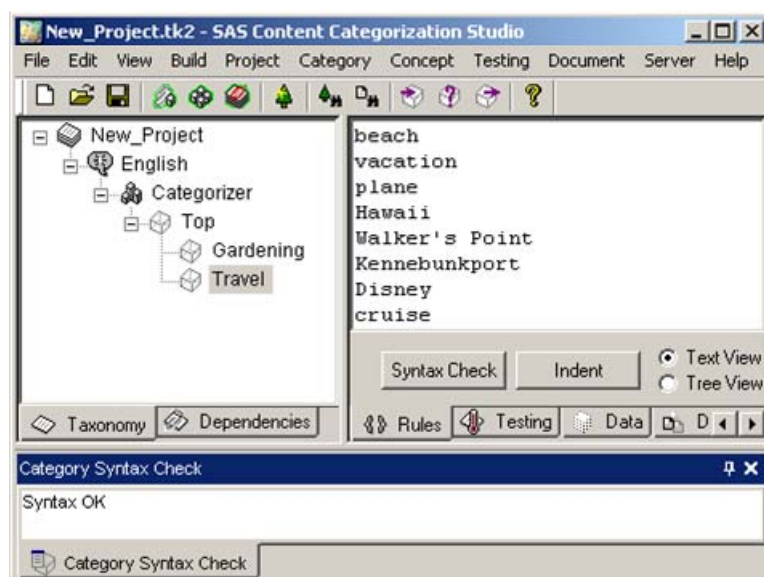
## 3.2 Writing Linguistic Rules

### 3.2.1 Write a List of Unique Identifiers

Linguistic rules, in their basic form, are lists of terms. Add special symbols or weights to increase their precision and recall.

To write a Linguistic rule, complete these steps:

1. Select **Project --> Settings**.
2. Under **Relevancy Type**, select **Frequency-Based** to match documents with the highest number of matching instances. Select **Zone-Based** to prioritize matches within certain document sections.
3. Type a list of unique identifying terms into the Rules window for the selected category. The default setting, **Text View**, is the only available selection for linguistic rules.



4. Click **Syntax Check**. The Category Syntax Check window appears at the bottom of the user interface with a status message. For example,

---

Syntax OK. If the syntax is not OK, the message contains information that enables you to make the necessary corrections.

5. (Optional) To increase the precision and recall of the rule, see one of the following sections:
  - Section 3.2.2 *Selecting Special Symbols* on page 25
  - Section 3.2.3 *Weighted Linguistic Rules* on page 27

### 3.2.2 Selecting Special Symbols

Special symbols for linguistic and Boolean rules differ. Some special symbols affect the match ratio setting and relevancy.

Table 3-1: Special Symbols Used in Linguistic Rules

Symbol	Type	Description
@	Suffix	Apply stemming to the word that precedes this symbol to expand the category rule so that it includes all forms of this word. For example, specify <code>price@</code> and the category rule expands to include <i>price</i> , <i>prices</i> , and <i>pricing</i> . The word, as well as all of its variants, count once if there is a match toward the match ratio. After the match ratio setting is met, each instance of a matching term and each stemming match count once toward frequency-based relevancy.
@N	Suffix	Expand the category to include all of the noun forms of the word that precedes this symbol. If the preceding word is not a noun, no stemming is applied. The word, as well as each of its matched variants, count once toward the match ratio specification and once toward frequency-based relevancy, after the match ratio is met.
@V	Suffix	Expand the category to include all of the verb forms of the word that precedes this symbol. If this term is not a verb, no stemming is applied. The word, as well as all of its variants, count once if there is a match toward the match ratio. Each word and stemming instance also count once toward frequency-based relevancy. This is true only after the match ratio is met.

Table 3-1: Special Symbols Used in Linguistic Rules

*	Prefix	Assign this term more classificatory weight (more relevancy) than other, unmarked words in the list. The single asterisk counts <i>twice</i> toward the match ratio, but only once toward relevancy.  This example uses a match ratio setting of 20%. If the term that is prefixed by * is matched, this term is worth 50% (10% of the 20% necessary) of the match ratio. It is then multiplied by 2, or 20%.
**	Prefix	Counts four times toward the match ratio, but only once toward relevancy.  Continue with the example of a match ratio setting of 20%. If the term that is prefixed by * is matched, the matching term is worth 50% (10% of the 20% necessary) of the match ratio multiplied by 4. 40% is double the 20% requirement for the match ratio setting.
-	Prefix	Counts against the match.
_L	Suffix	Use the underscore character (_) followed by an uppercase L to represent a literal. Append the at sign (@) to the end of a word, and the word is not expanded because it is treated as a literal.
_C	Suffix	Override case-insensitive using case-sensitive matching.
--	Prefix	Augment the single hyphen (-) symbol. The presence of these symbols causes the rule <i>not</i> to match. In other words, when this term is present and the match ratio setting is met, there is no category match for this document. (Frequency-based relevancy is irrelevant in this case.)
+	Prefix	Use this symbol to prevent a match if this term is not present in the document. This symbol also suppresses stemming, overrides the match ratio setting, and counts once toward frequency-based relevancy.
!	Suffix	Select <b>Expand all word forms</b> in the <b>Category</b> tab of the Project Settings window. All of the words in the category rule, except those that are followed by an exclamation point, are stemmed.

To weight your linguistic rules, complete these steps:

1. Qualify a linguistic rule with any of the special symbols that are described in the table above.

2. Click **Syntax Check** and the Category Syntax Check window appears at the bottom of the user interface. Use the messages that appear in this window to make changes to the rule.
3. Test your rules. For more information, see Chapter 4.

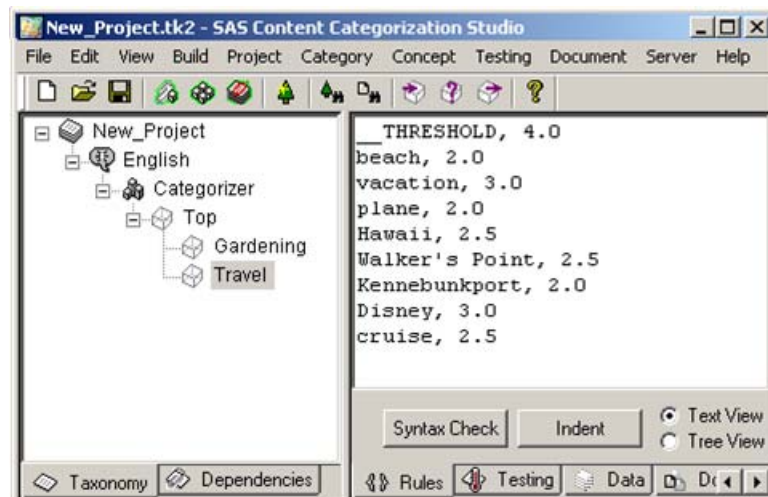
### 3.2.3 Weighted Linguistic Rules

Write weighted category rules for some or all of the categories in your taxonomy. In this case, weight is used to determine category membership. A weighted category rule specifies the weight assigned to each occurrence of a term. This rule also specifies the threshold that determines the sum of the weights necessary for category membership.

This type of category rule does not use any of the special symbols, relevancy, bias, or the match ratio specifications that are used with other forms of linguistic rules. When you weight a category, unless the rule terms occur with sufficient total frequency, the threshold weight is not met.

To add weights to linguistic rules, complete these steps:

1. Type a threshold weight into the first line of the category rule using the following syntax, `__THRESHOLD, <threshold_weight>`. (Do not type a space between the term, comma (,), or the weight.)



- 
2. Type a weight for each term using the following syntax:  
`<rule term>, <term_weight>.`
  3. Click **Syntax Check** and the Category Syntax Check window appears at the bottom of the user interface. Use the messages that appear to make any changes, if necessary, to the rule.
  4. Test your rules. For more information, see Chapter 4.

## 3.3 Writing Boolean Rules

### 3.3.1 Overview of Boolean Rules

Boolean rules differ from linguistic rules because the terms in Boolean rules are modified by Boolean terms. For this reason, you can begin the process of writing Boolean rules by writing lists of terms as if you are developing a linguistic rule. Add Boolean operators to these terms.

(Optional) To increase the precision and recall of the rule, see one of the following sections:

- Section 3.2.2 *Selecting Special Symbols* on page 25
- Section 3.2.3 *Weighted Linguistic Rules* on page 27

### 3.3.2 Adding Boolean Operators

See the following tables for the types and descriptions of the Boolean operators that you can add to your rules.



---

**Note:** Boolean operators are case sensitive.

---

Table 3-2: Common Boolean Operators

Operator	Description
AND	Takes two or more arguments. True only if all of the arguments are true.
OR	Takes two or more arguments. True if at least one argument is true.
NOT	Takes one argument. Use with the AND operator. True if the argument is false.
<b>Note:</b> The NOT operator is always used in the context of an AND, and not an OR operator.	
MIN_n	Takes one or more arguments. True if at least <i>n</i> arguments are true.
MINOC_n	Takes one or more arguments. True if the total number of occurrences of the arguments in the document is at least <i>n</i> .
MAXOC_n	Takes one or more arguments. True if the total number of occurrences of the arguments in the text is no more than <i>n</i> .
SENT	Takes two or more arguments. True if all of the arguments occur in the same sentence.
PAR	Takes two or more arguments. True if all the arguments occur in the same paragraph.
DIST_n	Takes two arguments. True if the first word appears in a sentence and the second word does not appear in the same sentence.
ORD	Takes two or more arguments. True if all of the arguments occur in the order specified by the rule.

---

The situational Boolean operators are listed below:

Table 3-3: Situational Boolean Operators

Operator	Description
NOTIN	Takes two arguments. True if the first argument occurs outside of the second argument. For example, use NOTIN, "health", "health care".
NOTDIST_n	Takes two arguments. True if both arguments are not within <i>n</i> words of each other.
NOTINSENT	Takes two arguments. True if both of the arguments appear in the same document, but not if they occur in the same sentence.
NOTINPAR	Takes two arguments. True if both of the arguments appear in the same document, but not if they occur in the same paragraph.
START_n	Takes one argument. True if the argument is matched within <i>n</i> words of the start of the document field.
END_n	Takes one argument. True if the argument is matched within <i>n</i> words from the end of the document.
ORDDIST_n	Takes two or more arguments. True if both arguments occur in the same order specified by the rule and if both occur within a distance of <i>n</i> words to each other.
MAXPAR_n	Takes one or more arguments. True if all arguments appear within the first <i>n</i> paragraphs.
MAXSENT_n	Takes one or more arguments. True if all arguments appear within the first <i>n</i> sentences.
PARPOS_n	Takes one or more arguments. True if all terms appear in the <i>n</i> <sup>th</sup> paragraph of the document.

---

### 3.3.3 Adding Special Symbols

Modify your Boolean rules using some of the special symbols that are also used for linguistic rules. These symbols are explained in the table below:

Table 3-4: Special Symbols Used in Boolean Rules

Symbol	Type	Description
@	Suffix	Use the at sign (@) to apply stemming to the word that precedes this symbol. The Boolean category rule is expanded to include all of its word forms. See the examples that follow this table.
@N	Suffix	Use the at sign (@) followed by N to expand the category rule to include all of the noun forms of the word that precede this symbol. For example, if you specify <code>book@N</code> , the category rule is expanded to include <code>books</code> . <b>Note:</b> If the preceding word is not a noun, no stemming is applied.
@V	Suffix	Use the at sign (@) followed by V to expand a word into all of the verb forms of the word. For example, if you specify <code>run@V</code> , the category rule is expanded to include <code>ran</code> , <code>run</code> , <code>running</code> , and <code>runs</code> . <b>Note:</b> If the preceding word is not a verb, no stemming is applied.
*	Suffix	Append the single asterisk (*), which is a wildcard character, to the end of a word. The asterisk matches any characters at the end of the word. For example (OR, "not*") matches <i>not</i> , <i>notebook</i> , <i>notice</i> , and <i>note</i> .
_L	Suffix	Use the underscore (_) and uppercase L together to match a literal. This combination matches a literal without the meaning associated with either of these special symbols. For example, see the following rules:  (OR, "end\$") match <i>end</i> at the end of the document.  (OR, "end\$_L") matches <i>end\$</i> , if it appears anywhere in the text.
_C	Suffix	Use the underscore (_) followed by the letter C to specify case-sensitive matching. For example (OR, "USA") matches <i>USA</i> , <i>usa</i> , <i>Usa</i> , and so on, while (OR, "USA_C") matches <i>USA</i> only.
_Q	Suffix	Use the underscore (_) followed by the uppercase Q to specify that any matching instances of this term qualify the document to match the rule. These matches do not contribute to the relevancy score for the document.

Table 3-4: Special Symbols Used in Boolean Rules (Continued)

<code>_C_Q</code>	Suffix	Use the suffix <code>_C_Q</code> (underscore [ <code>_</code> ] uppercase <code>C</code> followed by underscore uppercase <code>Q</code> ) after a word. These characters indicate that the qualifying, case-sensitive match does not contribute to the relevancy score.
<code>_L_Q</code>	Suffix	Use the suffix <code>_L_Q</code> (underscore [ <code>_</code> ] uppercase <code>L</code> followed by underscore uppercase <code>Q</code> ) after a term in a Boolean rule. This suffix qualifies a literal match. A match on this term makes the text a match for the category rule, but does not count when the relevancy score is computed.
<code>\$</code>	Suffix	Use the dollar sign ( <code>\$</code> ) to signal the end of a document. For example <code>(OR, "The End\$")</code> matches the string <i>The End</i> when the match occurs in the last string of the text. If the document contains the term <i>\$19.99</i> , this string can be matched as a literal. For example, this match is returned as a literal when <code>(OR, "\$19.99")</code> is specified.
<code>!</code>	Suffix	Use the exclamation point ( <code>!</code> ) to suppress stemming. If you select <b>Expand all word forms</b> in the <b>Category</b> tab, all of the words in the category rule, except those that are followed by an exclamation point, are stemmed.
<b>Note:</b> Most of these special symbols are also used for linguistic rules. For more information, see Section 3.2.2 <i>Selecting Special Symbols</i> on page 25.		

### 3.3.4 Expanding Word Forms

Click **Expand Forms** in the Rules window to see and test the list of terms that are possible rule matches when you append:

@

Expand this word into all of its word forms.

@N

Expand this word into all of its noun forms.

@V

Expand this word into all of its verb forms.

The expansion type that you specify for a term is automatically incorporated into the `<language>.mco` file. For this reason, you might want to return all expansions for the original form before you test your rules.

---

Click the **Expand Forms** button to see, and edit, the list of expansions that would otherwise automatically be applied to input documents for matching purposes. For example, a rule defining the *Safety* category might list *securities* as an expanded form of the word *security*. However, the word *securities* relates to financial markets and does not mean to be *protected* or *secure*. For this reason, if your rule specifies safety and protection, you should *not* append an @ sign to this term.

After you see and test the expanded word forms, you can select either **Expand words with '@' sign**, or **Expand all word forms** in the Project Settings - Category window.

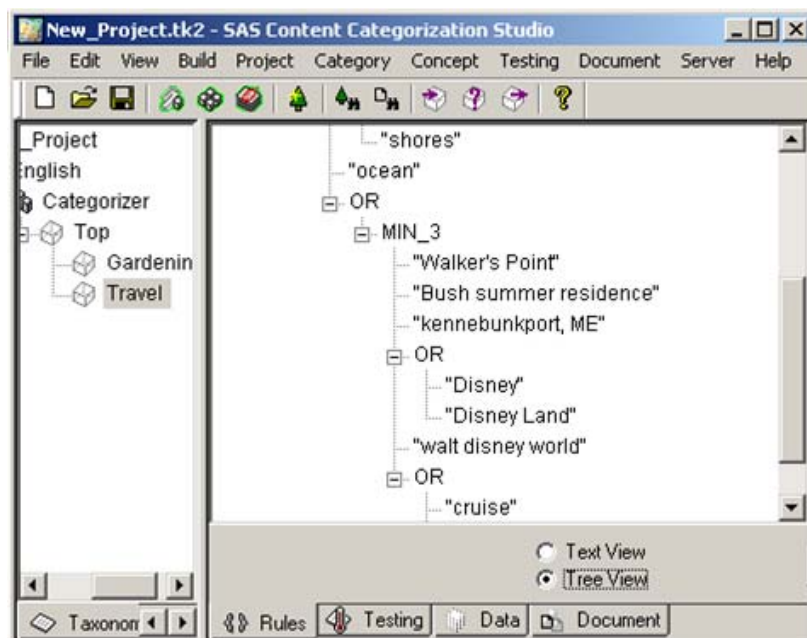
### 3.3.5 Specifying a Structured Text Field

If you are using Boolean rules to categorize Web documents, you can specify structured text fields and match attributes within these fields. For more information, see *SAS Content Categorization Studio: User's Guide*.

### 3.3.6 Viewing Boolean Rules

See a Boolean rule in tree view, or in text view. If you select **Text View**, you can also click **Indent** to see the rule as a taxonomy of terms and operators separated by parentheses ( ). Boolean operators separate the rule terms when you select **Tree View**.

Display 3-1 Boolean Rule Example



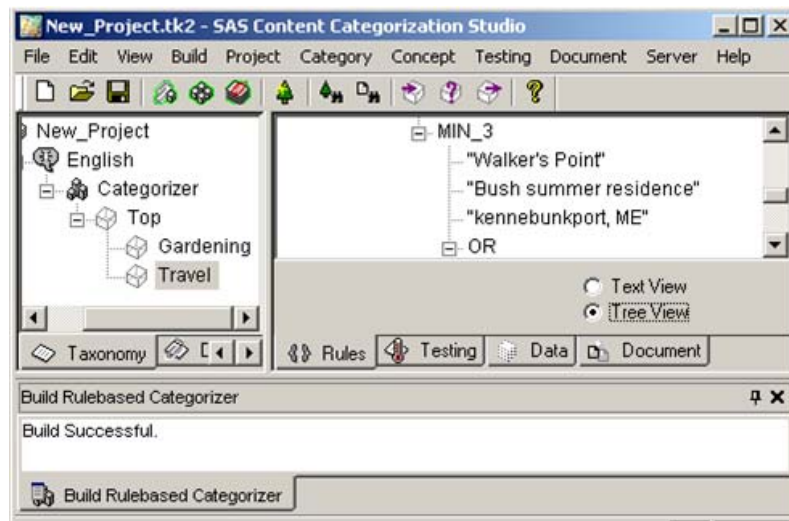
---

## 3.4 Build and Save the Categorizer

Build the categorizer to ensure that any changes that you make are applied to the input documents. Save the project to ensure that these changes are added.

To build the categorizer and save the project, complete these steps:

1. Select **Build --> Build Rulebased Categorizer** after you specify the rules for your categories.
2. See the status of the build. For example, see *Build Successful* in the Build Rulebased Categorizer window that appears at the bottom of the user interface.



3. Select **File --> Save**.

---

**Note:** If you select **Edit --> Options** and select **Always save before each test**, the Save operation is automatically performed when you select **TEST** in the **Testing** tab.

---

4. Use the build operation reiteratively every time you make changes to your project.





---

# 4

## Testing and Uploading Category Rules

---

- *Overview of Testing and Uploading Category Rules*
- *Testing Categories*
- *Use the Graphical Reports*
- *Test a Document in the Document Window*
- *Test against All Testing Documents*
- *Test Failing Test Files*
- *Upload the Categorizer to SAS Content Categorization Server*

### 4.1 Overview of Testing and Uploading Category Rules

After you create a taxonomy and write your category rules, you can test these rules to make sure that they work as expected. See the following paragraphs for an overview of the testing and uploading operations.

You gather groups of documents, or testing sets, for the purposes of testing the category rules that you develop in SAS Content Categorization Studio. These documents enable you to see the results that you can expect when SAS Content Categorization Server applies the rules to input texts.

You copy these sets of texts into a directory structure that mimics your taxonomy and you can also create a central repository of documents that should match. Although there are a number of testing operations that enable you to maximize the performance of your rules, this chapter covers only one set of directions.

---

Use this chapter to gain an overview of one way to test your categories. When you are satisfied with the results, upload the `.mco` file produced by SAS Content Categorization Studio to SAS Content Categorization Server.

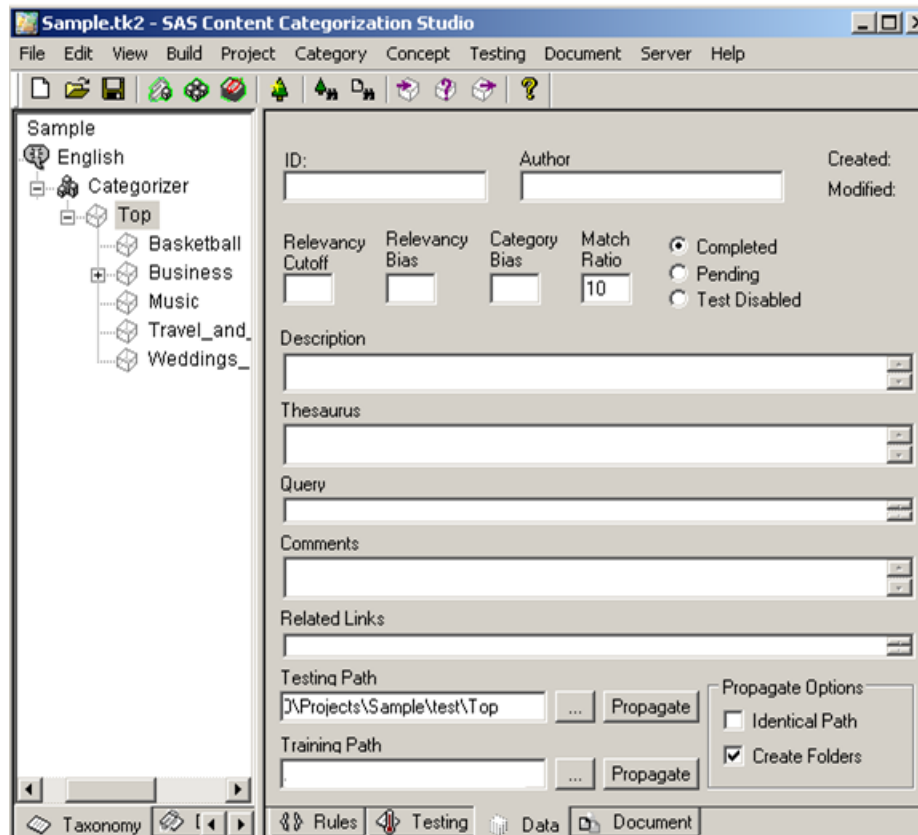
## 4.2 Testing Categories


### 4.2.1 Create a Testing Directory

Create a testing directory structure that holds all of the documents that you expect to match each category rule.

To create a testing directory, complete these steps:

1. Select the **Top** node in the Taxonomy window:



2. Select **Create Folders**.
3. Click  to the right of the **Testing Path** field and the Select a Directory window appears. Use this window to choose the location of your testing directory.
4. Click **Propagate** in the Data window and the testing paths map each category in your taxonomy to a testing folder in the testing directory.

---

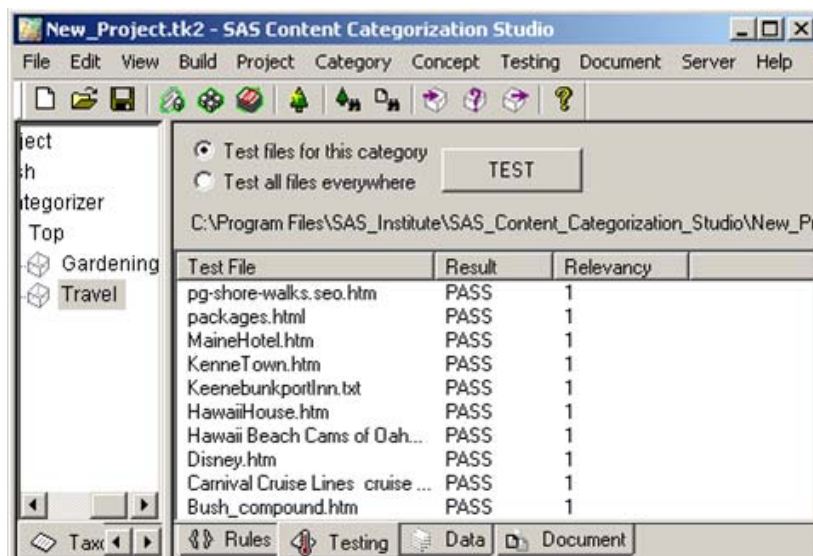
## 4.2.2 Populate Testing Directories

Assemble 10 or more documents for each category in the following formats .txt, .pdf, .xml, or .html texts. Use familiar texts that you expect to match the selected category rule. Place these testing files into the appropriate testing folders using cut and paste operations.

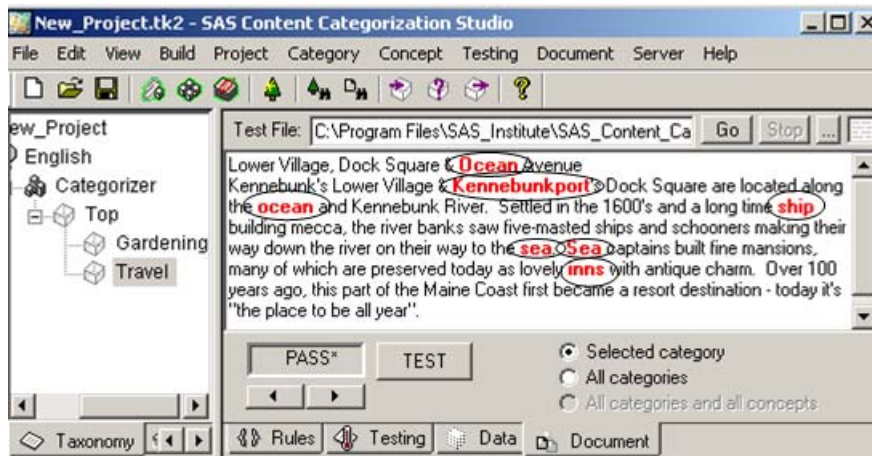
## 4.2.3 Test the Category Rule

To test your category rules, complete these steps:

1. Select a category in the Taxonomy window. For example, select `Travel`.
2. Click the **Testing** tab to see the testing files. By default the operation, **Test files for this category** is selected.
3. Click **TEST**. The testing results are displayed under the **Result** and **Relevancy** headings.



4. Double-click on a testing document to see the matched terms in the Document window. Matching terms are highlighted in red.



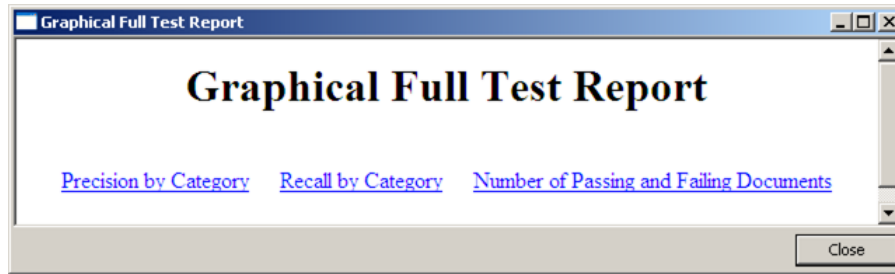
5. Redefine the tested category rule according to your testing results, reiteratively, if necessary. For example, adjust the **Default Relevancy Cutoff** setting in the Project Settings - Category window.

## 4.3 Use the Graphical Reports

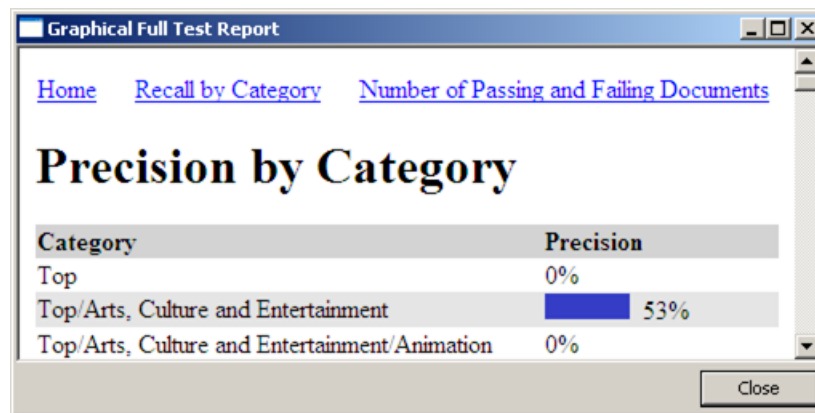
Use the graphical reports that are available in SAS Content Categorization Studio to see the precision, recall, and the numbers of passing and failing documents.

To open and use the Graphical full Test Report pages, complete these steps:

1. Select **Testing --> Graphical Full Test Report**. The Graphical Full Test Report page appears.

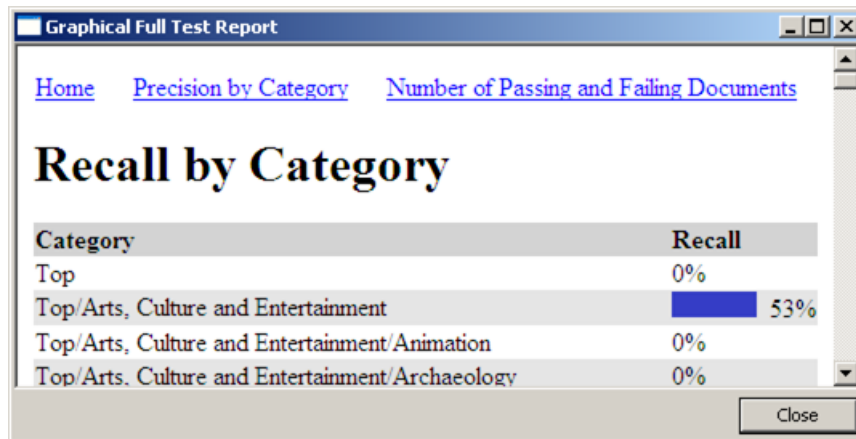


2. Click **Precision by Category**. The Precision by Category page appears.



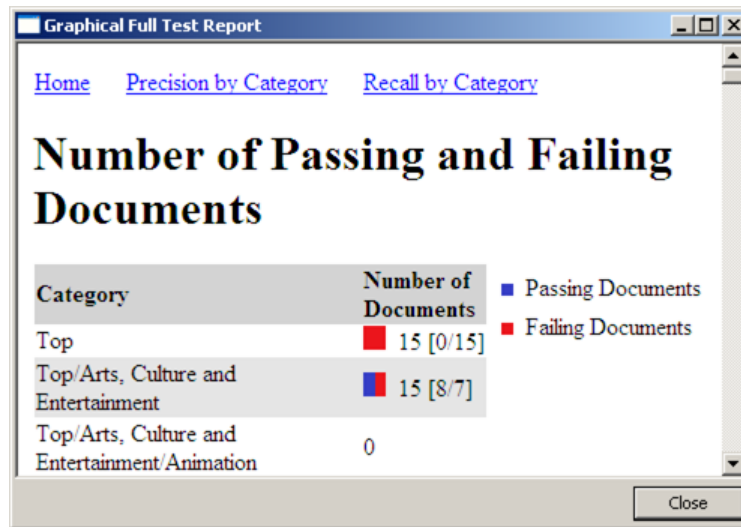
3. (Optional) Click the **Category** heading to display the categories alphabetically starting from the letter Z down or from the letter A down.
4. (Optional) Click the **Precision** heading to display the results starting from the 0%, or from 100%, down.

- 
5. Click **Recall by Category**. The Recall by Category page appears.



6. (Optional) Click the **Category** heading to display the categories alphabetically starting from the letter Z down or from the letter A down.
7. (Optional) Click the **Recall** heading to display the results starting from 0%, or from 100%, down.

- 
8. Click **Number of Passing and Failing Documents**. The Number of Passing and Failing Documents window appears.



9. See the number of **Passing Documents** in blue and the number of **Failing Documents** in red.
10. (Optional) Click the **Category** heading to display the categories alphabetically starting from the letter Z down or from the letter A down.
11. (Optional) Click the **Number of Documents** heading to display the results starting from the 0%, or from 100%, down.
12. Click **Close**.
13. (Optional) Click **Testing --> Show Last Full Graphical Testing Report** after you close this report. This operation restores the last report.



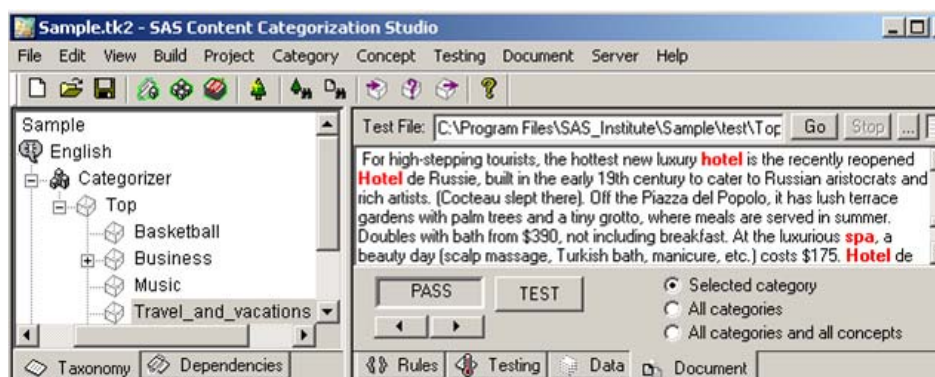
---



## 4.4 Test a Document in the Document Window

After you use the Testing window to test your documents, you can see the results for each text in the Document window. These results enable you to see the terms in your document that match the selected category rule. You can also test all of the categories in your taxonomy against the selected document. When you test all of your categories, you see all of the matching terms.

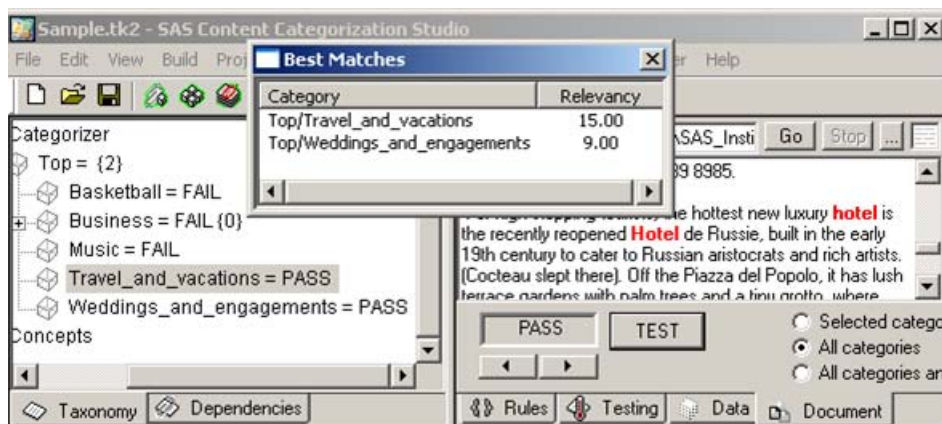
To test your documents in the Document window, complete these steps:

1. Double-click on a document in the Testing window and this text appears in the Document window.



2. By default you see the test results for the **Selected category** displayed in the document. Use the matching terms, highlighted in red, to see the words that made this document a passing text for the selected category.
3. Click the  and  to jump through each of the matches in the window.
4. (Optional) To remove the markup tags in an XML or HTML document, select **Document --> Remove Tags**. If you perform this operation, click **TEST** to see the tags reinstated.
5. A **PASS** or **FAIL** message for this text appears in the blank field to the left of the **TEST** button. Status messages are also displayed in the Taxonomy window, when the document is retested. These messages appear only if you select **All categories**.

The Best Matches window appears when you select **Edit --> Options** and select **Show best matches when testing all**. In addition, select **All categories**.



6. (Optional) Use the Best Matches window to see the matching nodes under the **Category** heading. You can also see the relevancy score for each passing category under the **Relevancy** heading.

## 4.5 Test against All Testing Documents

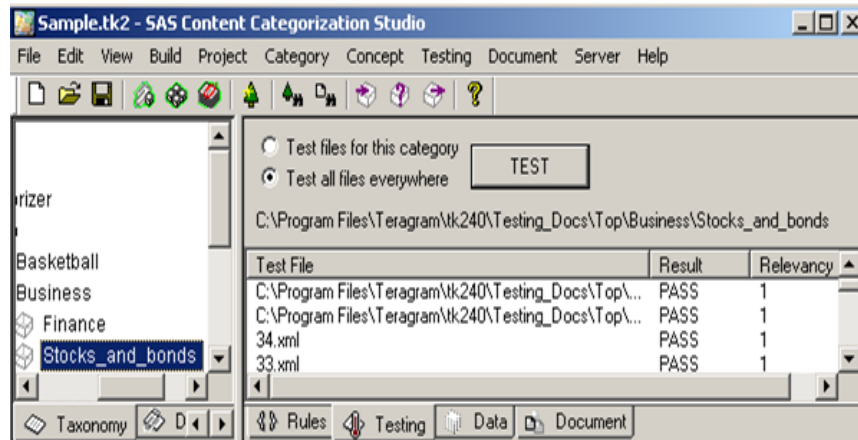
You can choose to test all of the documents in the testing folders for all of the categories in the taxonomy. When you choose to perform this operation, you can see documents that match this category, but which are selected to match another node.

After you perform this operation, you can access the tested document in the Document window. Use the Document window to compare the matches for two or more categories.

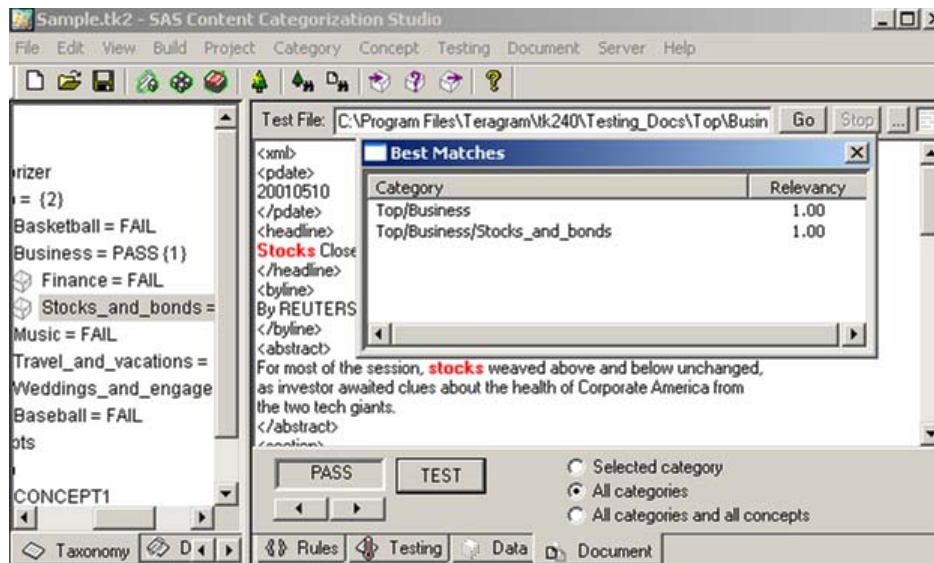
To test all of your testing document, complete these steps:

1. Select **Test all files everywhere** in the **Testing** tab.

2. Click **TEST**.



3. Double-click a document that passes and is out-of-category. The document appears in the Document window.



4. Select **All categories**.

5. Click **TEST**.

6. Examine and compare the matches that are highlighted in red.

---

## 4.6 Test Failing Test Files

After you refine a rule, select files that passed, but should have failed the testing operation. For example, assign failing status to *cruise ship* documents for the *Tom Cruise* set of testing documents. Move this file, and any others, to the directory that you set up for failing files. You can use this directory at the end of the testing process to ensure the accuracy of your rules.

To import failing documents complete these steps:

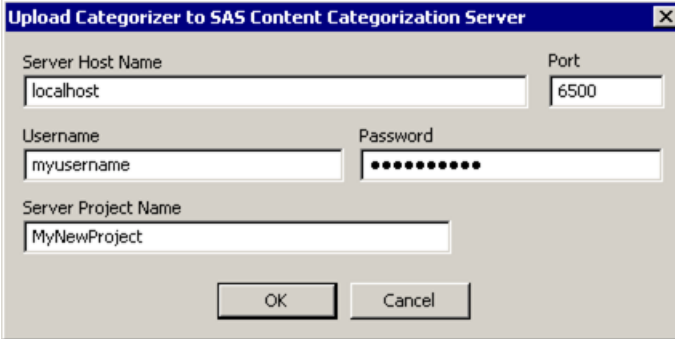
1. Click the **Testing** tab.
2. Select **Testing --> Import Failing Test Files**.
3. Click **TEST**.
4. Double-click any failing file that passes to see the results in the **Document** tab.

## 4.7 Upload the Categorizer to SAS Content Categorization Server

After you build and test the taxonomy, use the upload categorizer to SAS Content Categorization Server operation. You can specify the requirements that are necessary to upload the categorizer (.mco file) to the server in this window.

To upload the .mco file to SAS Content Categorization Server, complete these steps:

- 
1. Select **Build --> Upload Categorizer**. The **Upload Categorizer to SAS Content Categorization Server** window appears.



The screenshot shows a dialog box titled "Upload Categorizer to SAS Content Categorization Server". It contains four input fields: "Server Host Name" with the value "localhost", "Port" with the value "6500", "Username" with the value "myusername", and "Password" with a masked password represented by ten dots. Below these fields is a "Server Project Name" field with the value "MyNewProject". At the bottom of the dialog are "OK" and "Cancel" buttons.

2. By default, the **Server Host Name** field is entered for you. If necessary, enter a new server name.
3. By default, the **Port** field is entered for you. If necessary, enter a new port number.
4. Enter your user name into the **Username** field.
5. Enter your password into the **Password** field.
6. Enter the name of your project into the **Server Project Name** field.
7. Click **OK**.
8. Begin applying the rules to input documents using SAS Content Categorization Server. For more information, see *SAS Content Categorization Server: User's Guide*.



---

# Appendixes

---

- Appendix A: *Recommended Reading on page 53*
- Appendix B: *Glossary on page 55*





---

# Appendix: A

## Recommended Reading

---

The following books are recommended:

- *SAS Content Categorization Studio: User's Guide*: Create a SAS Content Categorization Studio project, test, and upload to SAS Content Categorization Server.
- *SAS Content Categorization Studio: Installation Guide*: Install SAS Content Categorization Studio.
- *SAS Contextual Extraction Studio: User's Guide*: Use this add-on application to SAS Content Categorization Studio to write complex concept definitions that can include multiple rule types within a single definition.
- *SAS Contextual Extraction Studio: Installation Guide*: Install SAS Contextual Extraction Studio.
- *SAS Content Categorization Collaborative Server: User's Guide*: Use this add-on application to SAS Content Categorization Studio to enable multiple users to build a single project.
- *SAS Content Categorization Server: Administrator's Guide*: Understand how SAS Content Categorization Server applies the binary files output by SAS Content Categorization Studio to input documents.
- Use the language books for each language purchased to see the comprehensive list of part-of-speech tags that are available for grammar concepts.

SAS offers instructor-led training and self-paced e-learning courses to help you get started with the SAS add-in, learn how the SAS add-in works with the other products in the SAS Enterprise Intelligence Platform, and learn how to run stored processes in the SAS add-in. For more information about the courses available, see [support.sas.com/training](http://support.sas.com/training).

For a complete list of SAS publications, see the current SAS Publishing Catalog. To order the most current publications or to receive a free copy of the catalog, contact a SAS representative at

---

SAS Publishing Sales  
SAS Campus Drive  
Cary, NC 27513  
Telephone: (800) 727-3228\*  
Fax: (919) 677-8166  
E-mail: [sasbook@sas.com](mailto:sasbook@sas.com)  
Web address: [support.sas.com/pubs](http://support.sas.com/pubs)  
\* For other SAS Institute business, call (919) 677-8000.

Customers outside the United States should contact their local SAS office.

---

# Appendix: B

## Glossary

---

### **Batch testing**

process of testing all of the testing documents in the testing set against a selected category. Alternatively, choose to test all of the documents in the testing set against the entire taxonomy. In the second case, use testing documents that were not specifically selected for a category to gain comprehensive testing results that simulate real usage.

### **Categorization**

process of concisely defining the subject matter of a document, in other words, the main idea or subject of the document.

### **Central repository of documents**

place all of your testing documents into one directory instead of creating a directory structure that matches your taxonomy structure. Alternatively, you can use a central repository as a secondary source for testing documents.

### **Frequency-based ranking**

is the number of matching terms that are found in a document.

### **Hierarchical taxonomy**

build a taxonomy with subcategories and possibly subcategories or children of these child nodes in the taxonomy tree.

### **Linguistic rules**

specify the key words that define the categories in the taxonomy.

### **Match Ratio**

specify the percentage of terms in the category rule to be matched.

### **Metadata**

is data on information.

### **Precision**

measure the relevancy of the matched documents. In other words, the category rule excludes possible matches that do not reflect the subject

---

---

matter of the category. For example, texts that refer to *rock collections* are not matched for the category *Rock and Roll*.

**Recall**

measure whether all of the relevant texts are matched by the category rule.

**Relevancy-biased ranking**

is the measure of the appropriateness of the match for one category within the overall taxonomy.

**Relevancy range**

specify the appropriateness of the documents to a specific category.

**String**

is a group of words or characters that you specify for a rule.

**Structured text field names**

specify searchable field names in an HTML, SGML, or XML document for a rule-based categorizer using Boolean terms.

**Taxonomy**

organize a classification structure that can be either a flat or a hierarchical system.

**Testing Set of documents**

is the set of texts that you use to test the categorizer or concepts extractor.

**Threshold**

specify the minimum weight that is necessary to be considered a member of the selected category when a weighted linguistic rule is specified. This numerical threshold is specified in the Rules window using `__Threshold`, followed by the number to be matched or exceeded for each specified term. If the total weight of the occurrences of terms in a selected document equal or exceed this number, the document might be a match for this category.

**Weight**

specify a number that equals the weight assigned to each term in a weighted category rule. Assign the most important terms higher weights than those of less significance.

---

# Index

---

-	special symbol .....	26
--	category rule .....	26
	special symbol .....	26
!	category rule .....	26
	special symbol .....	26, 32
\$	special symbol .....	32
*	special symbol .....	26, 31
**	special symbol .....	26
+	category rule .....	26
	special symbol .....	26
.mco file	usage .....	6
@	special symbol .....	25, 31
@N	special symbol .....	25, 31
@V	special symbol .....	25, 31
_C	special symbol .....	26, 31
_C_Q	special symbol .....	32
_L	special symbol .....	26, 31
_L_Q	special symbol .....	32
_Q	special symbol .....	31

---

---

## A

Add Category	
usage .....	20
Add Language	
usage .....	10
All categories	
usage .....	45
Allow Duplicate ID's	
defined .....	16
Allow Short Macro Names	
defined .....	15
Always rebuild before each test	
usage .....	12
Always save before each test	
usage .....	12
AND operator .....	29
Auto-Rule Generation Max Words	
defined .....	17

## B

Boolean Morphological Expansion	
defined .....	16
Boolean rules	
defined .....	23
example .....	34
usage .....	23
Web document .....	33
Build	
Build Rulebased Categorizer .....	35
build	
categorizer .....	8
status .....	35
Build Rulebased Categorizer	
Build option .....	35
Build Rule-based Categorizer window	
open .....	35

---

## C

categorization	
defined .....	7
categorizer	
build .....	8
category name	
enter .....	20
category rules	
test .....	40
Category Syntax Check window	
usage .....	24
Check info strings for duplicates	
usage .....	13
Check match strings for duplicates	
usage .....	13
Compatibility Date	
Misc window .....	18
concept definition	
precision .....	7
recall .....	7
Custom Syntax Checker Executable	
Misc window .....	19

## D

Default Category Bias	
defined .....	15
Default Relevancy Cutoff	
defined .....	15
Disable Substring Matches	
defined .....	16
DIST_n operator .....	29
Document window	
usage .....	41

---

## E

Enable Categorizer	
usage .....	11
END_n operator .....	30
Expand all word forms	
defined .....	17
Expand word forms with '@' sign	
defined .....	17
Export MCO file with UTF-8 Display Names	
defined .....	16
Export Short MCO file	
defined .....	16

## F

Flag categories/concepts with no definitions	
Options window .....	13
Flag categories/concepts with no dependencies	
Options window .....	13
Frequency-Based	
defined .....	15

## H

Hide Display Names for UTF-8 Languages option	
usage .....	13

## I

Individual Field Anchors	
Misc window .....	19

## L

Linguistic rules	
defined .....	23, 24
options .....	23
text view .....	24



---

## M

match ratio	
special symbols .....	25
MAXOC_n operator .....	29
MAXPAR_n operator .....	30
MAXSENT_n operator .....	30
MIN_n operator .....	29
MINOC_n operator .....	29
Misc tab	
XML Tags to Ignore .....	19
Misc window	
Compatibility Date .....	18
Custom syntax Checker .....	19
Individual Field Anchors .....	19
Paragraph Separator .....	19
Use UTF-8 Test Files .....	18
XML Default Field .....	19

## N

Never expand word forms	
defined .....	16
new project	
create .....	9
NOT operator .....	29
NOTDIST_n operator .....	30
NOTIN operator .....	30
NOTINPAR operator .....	30
NOTINSENT operator .....	30
noun form	
category rule .....	25, 31

## O

Operator-Based	
defined .....	15
Optimize for	
defined .....	17
OR operator .....	29

---

ORD operator .....	29
ORDDIST_n operator .....	30

## P

PAR operator .....	29
Paragraph Separator	
Misc window .....	19
PARPOS_n operator .....	30
precision	
adequacy .....	8
concept definition .....	7
define .....	5
Project Location	
New Project window .....	9
Project Name	
New Project window .....	9
Project Settings windows	
usage .....	14
Propagate button	
usage .....	39

## R

recall	
adequacy .....	8
concept definition .....	7
define .....	5
Relevancy Type	
defined .....	15
Report duplicate entries when checking classifier concepts	
usage .....	13
rules	
write .....	8

---

## S

Select a Directory window	
usage .....	39
Select a Language window	
open .....	10
SENT operator .....	29
Show best matches when testing all	
Test all files everywhere .....	12
usage .....	12
Sort taxonomies automatically	
option .....	13
special symbol	
- .....	26
-- .....	26
! .....	26, 32
\$ .....	32
* .....	26, 31
** .....	26
+ .....	26
@ .....	25, 31
@N .....	25, 31
@V .....	25, 31
_C .....	26, 31
_C_Q .....	32
_L .....	26, 31
_L_Q .....	32
_Q .....	31
special symbols	
match ratio .....	25
START_n operator .....	30
stemming	
category rule .....	25, 31
Syntax Check button	
usage .....	24

---

## T

Taxonomy as Text	
window .....	13
Test all files everywhere	
Show best matches when testing all .....	12
TEST button	
usage .....	40
Test files for this category	
usage .....	40
testing directory	
create .....	38
defined .....	38
testing files	
fail .....	8
pass .....	8
Testing Path field	
usage .....	39
Testing window	
usage .....	40
Text View	
Linguistic rules .....	24
threshold weight	
total weight .....	27
total weight	
threshold weight .....	27

## U

Use UTF-8 Test Files	
Misc window .....	18

## V

verb form	
category rule .....	25, 31

---

## W

Web document	
Boolean rules .....	33
weight	
weighted category rule .....	27
weighted category rule	
weight .....	27

## X

XML Default Field	
Misc window .....	19
XML Tags to Ignore	
Misc tab .....	19

## Z

Zone-Based	
defined .....	15

