



THE
POWER
TO KNOW.

SAS® Content Categorization Single User Servers 12.1 Administrator's Guide



The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2012.
SAS® Content Categorization Single User Servers 12.1: Administrator's Guide. Cary, NC:
SAS Institute Inc.

SAS® Content Categorization Single User Servers 12.1: Administrator's Guide

Copyright © 2012, 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, March 2012

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

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

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

Contents

About This Book	1
Audience	1
Prerequisites	1
Conventions	2
 What's New in SAS Content Categorization Single User Servers 12.1	 3
 1 About SAS Content Categorization Single User Servers	 5
1.1 What is SAS Content Categorization Single User Servers?	5
1.2 Benefits of Using SAS Content Categorization Single User Servers	6
1.3 How Does SAS Content Categorization Single User Servers Work?	7
1.4 About the Architecture	7
 Part 1: Installing the Servers	 9
 2 Using the Installation and Related Processes	 11
2.1 Before You Install the Servers	11
2.1.1 Overview of Installation	11
2.1.2 Prerequisites	12
2.1.3 Using the SAS License File	13
2.2 Installing the SAS Content Categorization Single User Servers	14
2.2.1 Install on Windows	14
2.2.1.A Before You Install on Windows	14
2.2.1.B Installing on Windows	14
2.2.1.C The Folders That Appear after Installation	27
2.2.2 Install on UNIX	27
2.2.2.A Install on UNIX	27
2.2.2.B See the Directory Structure	28
2.3 Starting and Stopping a Server on a Windows Machine	29
2.4 Uninstall on Windows	30
2.5 Access the Servers	33
2.6 Processing Documents	34

Part 2: SAS Content Categorization Server 35

3 Configuring and Running SAS Content Categorization Server37

3.1 Configuration Overview	37
3.2 SAS Content Categorization Server Configuration File	39
3.2.1 Windows Configuration File	39
3.2.2 UNIX Configuration File	39
3.2.3 The Directives	40
3.3 Specifying Project Users and Creators	44
3.4 Add a Project	45
3.5 Using the Models Directory	48
3.6 Using the Descriptors Directory	49
3.7 Using cat_log and concept_log Files	50
3.8 Specifying Multiple Project Files	51
3.9 Sending Documents to the Server	52
3.10 Run SAS Content Categorization Server	52
3.10.1 Verify That the Server Is Running	52
3.10.2 Run the Server on UNIX	53
3.11 Optimize Performance on a Client Windows Machine	53
3.11.1 Overview of Performance Optimization	53
3.11.2 Before and After You Optimize Performance	53
3.11.3 Adjust the TCP Time Wait State	54
3.11.4 Reset Ephemeral Ports	54

4 SAS Content Categorization Server Web Administration57

4.1 Overview of SAS Content Categorization Server Web Administration	57
4.2 Access the Administration Web Interface	58
4.3 Using the Administration Web Page	59
4.3.1 Overview of the Administration Web Page	59
4.3.2 Use the Are you there? Page	59
4.3.3 Use the SAS Content Categorization Server Projects List Page	60
4.3.4 Use the SAS Content Categorization Server Categorization Statistics (Matches) Page	63
4.3.5 Use the SAS Content Categorization Server Categorization Statistics (Timing) Page	64
4.3.6 Use the SAS Content Categorization Server Concept Extraction Statistics Page	67
4.3.7 Use the SAS Content Categorization Single User Servers Concept Extraction Statistics (Timing) Page	69

Part 3: SAS Document Conversion Server	73
5 Processing Various File Types into Plain Text Format	75
5.1 Overview of SAS Document Conversion Server	75
5.2 The File Types That SAS Document Conversion Server Can Convert	76
Part 4: Appendixes	77
A Troubleshooting	79
A.1 Installing SAS Content Categorization Single User Servers	79
A.1.1 The Installer Fails to Copy Files on Windows Vista or Server 2008	79
A.2 Tips and Guidelines for SAS Content Categorization Server	80
A.2.1 If You Are Unable to Upload a File to SAS Content Categorization Server	80
A.2.2 Trying to Upload a Project with a Duplicate Name	80
A.2.3 Noun Phrases Incorrectly Split by an Apostrophe (‘)	80
A.2.4 Using Synonym Lists	80
A.3 Using the Configuration File	81
A.3.1 Trailing Spaces in the Configuration File	81
A.3.2 Modifying the Configuration Backup File	81
A.4 If SAS Content Categorization Server Does Not Appear to Be Running ..	82
A.4.1 Overview of When the Server Does Not Appear to Be Running ...	82
A.4.2 Checking and Debugging on a Windows Machine	82
A.4.3 Checking and Debugging on a UNIX Machine	83
B Recommended Reading	85
C Glossary	87
Index	89

About This Book

Audience

SAS Content Categorization Single User Servers is designed for the following types of administrators and users, depending on the server components that you install:

- Users who want to run SAS Document Conversion Server to preprocess files into plain text format.
- Administrators and users who upload `.concepts` or `.mco` binary files from a SAS Content Categorization Studio project to SAS Content Categorization Server.
- Users who access the Administration Web Page to see the matching statistics for categories and concepts that are applied to input documents by SAS Content Categorization Server.
- Programmers who want to use the supplied APIs to write their own applications.

Prerequisites

Here are the prerequisites for using SAS Content Categorization Single User Servers:

- Install Java 1.5 or higher and Python version 2.3 or higher if you plan to install the Java and Python APIs for SAS Content Categorization Server. If you plan to install SAS Document Conversion Server Java API, make sure that the Java 1.5, or higher, run-time environment is installed on the machine.
- Install one or more of the following SAS Content Categorization Single User Servers components:
 - SAS Document Conversion Server: Preprocess your documents into plain text format.

-
- SAS Content Categorization Server: Apply the categories or concepts that you uploaded to this server to input documents.
 - Configure the server, or servers, that you install.

Conventions

This manual uses parts to define the chapters that are install or component specific. This manual also uses the following typographical conventions:

Convention	Description
TGM_ROOT	The root directory where SAS Content Categorization Single User Servers is installed, typically the following: Windows: C:\Program Files\Teragram\SAS Content Categorization Single User Servers\ UNIX: /opt/sas_catcon_servers_linux64
Next button	The labels for user interface controls are shown in a bold, sans-serif font.
<u>www.sas.com</u>	The hypertext links are shown in a light blue, fixed-width font, and are underlined.

What's New in SAS Content Categorization Single User Servers 12.1

The new features in SAS Content Categorization Single User Servers include the following:

- The addition of SAS Document Conversion Server to the SAS Content Categorization Server package.
- A single installer for the package.
- Python Web services API.

Chapter: 1

About SAS Content Categorization Single User Servers

- *What is SAS Content Categorization Single User Servers?*
- *Benefits of Using SAS Content Categorization Single User Servers*
- *How Does SAS Content Categorization Single User Servers Work?*
- *About the Architecture*

1.1 What is SAS Content Categorization Single User Servers?

In most organizations it is necessary to obtain information about, and from, data that is created internally and externally. SAS Content Categorization Single User Servers installs several applications to perform these tasks. You can choose to install any, or all, of the servers included in SAS Content Categorization Single User Servers:

- Use SAS Document Conversion Server to preprocess your documents, converting them from various file types into plain text.
- Use SAS Content Categorization Server to automatically apply categorization and concept extraction to input documents. Categorization and concept extraction are applied using the category rules and concept definitions that your organization develops in SAS Content Categorization Studio.
- Write scripts to preprocess documents and apply categorization and concept extraction to input documents using Java and Python APIs.

The installation and configuration documentation for each server is found in this book. Some applications also have their own documentation that can be found in the **Help** menu or in the product name folder.

Easy document preprocessing

Use SAS Document Conversion Server to preprocess your files into plain text format.

Easy configuration

Most of the configuration file for SAS Content Categorization Server is written for you, and uploaded binary files automatically appear in the configuration file.

Easy monitoring

You can use the Administration Web Page to see whether SAS Content Categorization Server is running. You can also see data on concept and category matches, input documents, and timing.

1.2 Benefits of Using SAS Content Categorization Single User Servers

SAS Content Categorization Single User Servers provides users with the following benefits:

Convert documents into plain text

SAS Document Conversion Server automatically converts input documents, of various types, into plain text.

Automatically locate matching documents

SAS Content Categorization Server automatically applies the category rules and concept definitions developed in SAS Content Categorization Studio to input documents.

Gain real-time knowledge of matches

The Administration Web Page enables you to see the statistics generated by matching concepts and categories in real time.

Save money on information retrieval and organization costs

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

Write custom programs

Developers use the included APIs to write applications to perform document conversion.

1.3 How Does SAS Content Categorization Single User Servers Work?

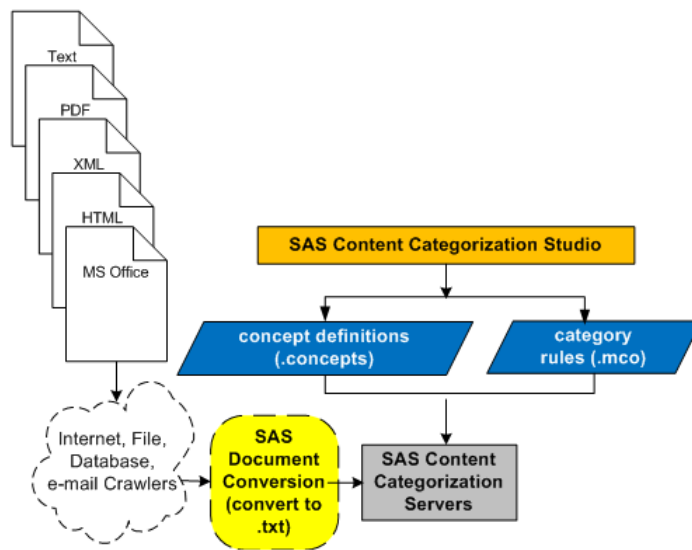
SAS Content Categorization Single User Servers consists of several applications. Administrators and other users use SAS Document Conversion Server to automatically convert documents to plain text format. These users upload models and input documents to SAS Content Categorization Server where automatic document categorization and concept extraction is performed. SAS Content Categorization Single User Servers applies the rules and definitions in the form of binary files. These rules and definitions are written in SAS Content Categorization Studio.

1.4 About the Architecture

SAS Content Categorization Single User Servers provides rapid, run-time document conversion, categorization, and concept extraction for documents collected from your corporate intranet or the Internet. The SAS Document Conversion Server converts documents into plain text that can be used by SAS Content Categorization Server. SAS Content Categorization Server automates the application of the compiled category rules and concept definitions in the binary files that are uploaded from SAS Content Categorization Studio.

You can also use the Administration Web Page to see various types of statistical reports on the matched categories and concepts.

Figure 1-1 SAS Content Categorization Single User Servers Architecture



Part 1: Installing the Servers

- Chapter 2: *Using the Installation and Related Processes on page 11*

Chapter: 2

Using the Installation and Related Processes

- *Before You Install the Servers*
- *Installing the SAS Content Categorization Single User Servers*
- *Starting and Stopping a Server on a Windows Machine*
- *Uninstall on Windows*
- *Access the Servers*
- *Processing Documents*

2.1 Before You Install the Servers

2.1.1 Overview of Installation

This chapter explains the hardware requirements and the installation process for SAS Content Categorization Single User Servers. This chapter also explains how to specify the path to the SAS license that is necessary for installation.

The SAS Content Categorization Single User Servers installation kit for Windows contains all of the components required to install (and uninstall) SAS Content Categorization Single User Servers. For example, use `SAS_ConCat_Servers_Win32_Setup.exe` or `SAS_ConCat_Servers_Win64_Setup.exe` to install on Windows. These servers are:

- SAS Content Categorization Server
- SAS Document Conversion Server

and the necessary APIs.

The installation is performed by a system administrator who is familiar with the operating system and who has sufficient system privileges to create directories and to define user permissions.

In some cases, you might want to install, or uninstall, one or more of the components of SAS Content Categorization Single User Servers. For example, you might want to remove a local copy of SAS Content Categorization Server from your machine because you are connecting to the main server.

To perform this operation, complete these steps:

1. Install SAS Content Categorization Server and the SAS Content Categorization Java API when you install SAS Content Categorization Single User Servers.
2. Use SAS Content Categorization Server to verify that the Java API code is written correctly.
3. Use the Java code to connect to the main SAS Content Categorization Server.
4. Uninstall the local copy of SAS Content Categorization Server taking care not to uninstall SAS Content Categorization Java API. (For more information, see Section 2.4 *Uninstall on Windows* on page 30.)

2.1.2 Prerequisites

Configure the machine where you install SAS Content Categorization Single User Servers according to the recommended system configuration:

CPU

x86 with 1 GHz or higher required. 2+ CPUs of 2 GHz or higher, each, are recommended

RAM

1 GB or higher is recommended, but this base number depends on the number of binary files that you load

The table below lists the hardware requirements that are necessary to run SAS Content Categorization Single User Servers:

Table 2-1: Supported Operating Systems

Operating System	Platform
Linux, (Red Hat 7.x, 8, 9, Fedora 1-3, RHEL 2.1 and higher), SUSE	x86, x86-64
IBM AIX	PPC
HP-UX	PA-RISC
Sun Solaris (32-bit)	SPARC
Sun Solaris (64-bit)	UltraSPARC, x86-64
Windows	x86, x86-64

Note: Ensure that Java run-time environment 1.5 or later is installed on your machine or you cannot install SAS Document Conversion Server and Java API.

2.1.3 Using the SAS License File

This SAS license is the SAS installation data file (SID file) that is included in the Software Order E-mail (SOE) that you received. Save the setinit file `tg-master-noecc.sas` to a directory on your hard drive. When you locate this file during installation, the folder location and path are written to the `server.config` file located in the Teragram CatCon Server/conf folder. For more information, see Section 3.2 *SAS Content Categorization Server Configuration File* on page 39. If the path to this file does not appear in your server configuration file, see Section A.3.2 *Modifying the Configuration Backup File* on page 81.

2.2 Installing the SAS Content Categorization Single User Servers

2.2.1 Install on Windows

2.2.1.A Before You Install on Windows

When you work through the SAS Content Categorization Single User Servers Setup wizard, there are some steps that pertain only to SAS Content Categorization Server. These pages specify SAS Content Categorization Server in the directions provided.

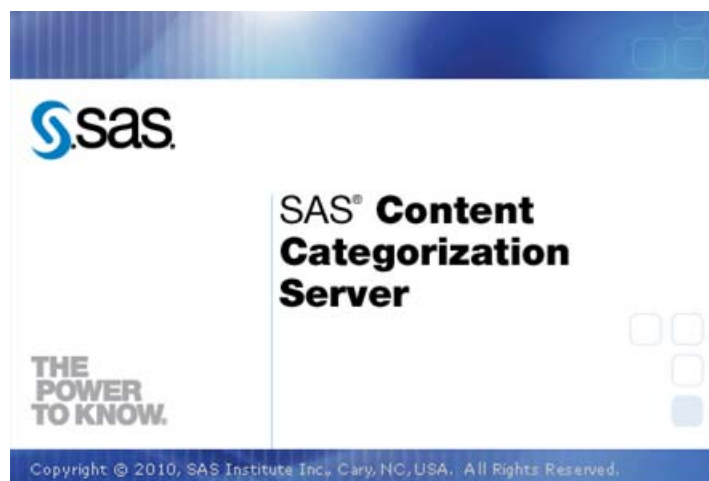
You should install Java 1.5 or higher and Python version 2.3 or higher if you plan to install the Java and Python APIs for SAS Content Categorization Server. If you plan to install SAS Document Conversion Server and Java API, make sure that the Java 1.5, or higher, run-time environment is installed.

Notes: SAS Document Conversion Server and Java API cannot be installed unless the correct Java run-time environment is installed.

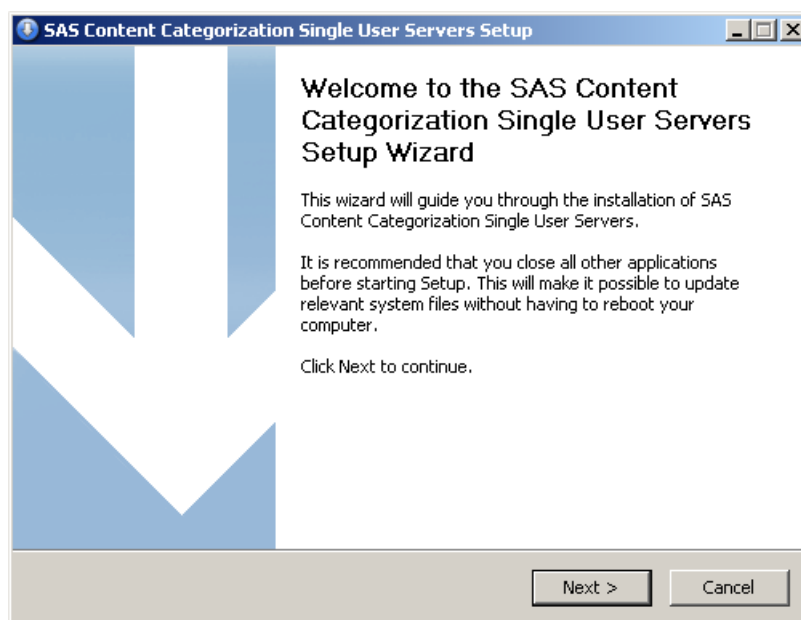
2.2.1.B Installing on Windows

To install the SAS Content Categorization Single User Servers software on a supported Microsoft Windows system, complete these steps:

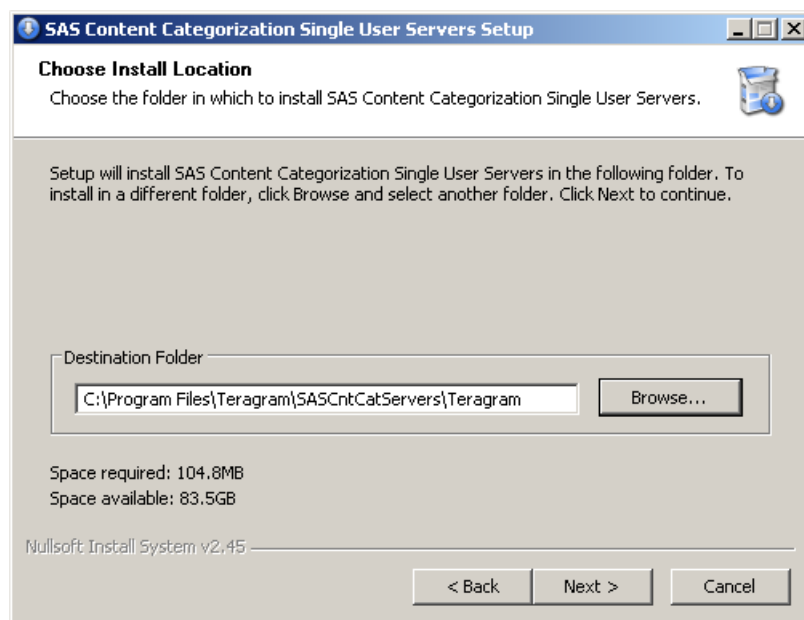
1. Double-click `SAS_ConCat_Servers_<arch>_Setup.exe` and the installation wizard appears.



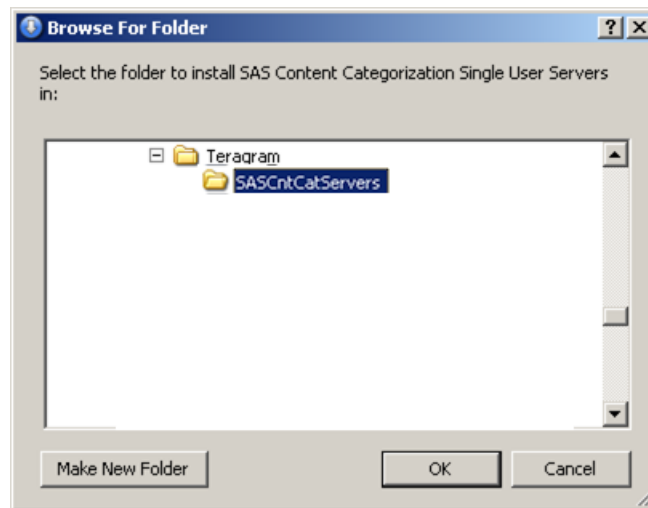
The Welcome page appears.



-
2. Click **Next** and the Choose Install Location page appears where you can enter the path to the installation folder.

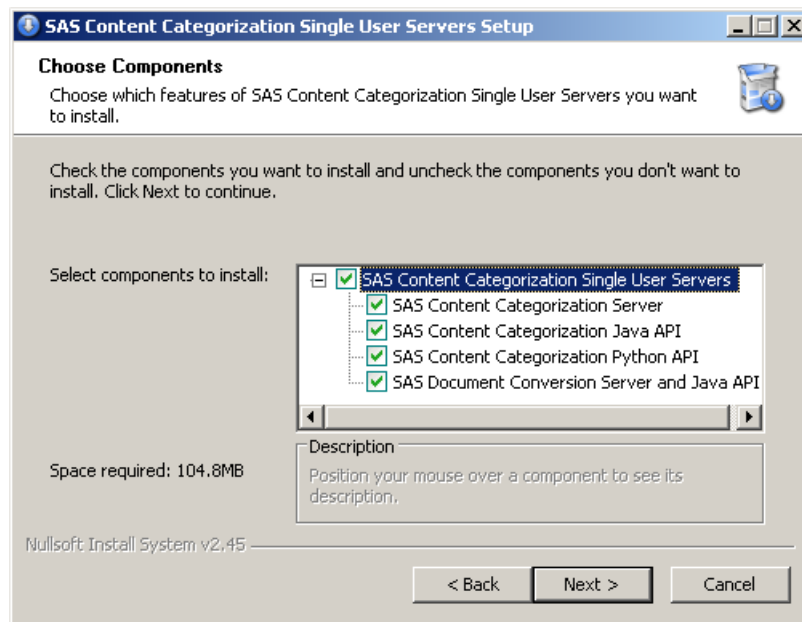



-
3. (Optional) Click **Browse** and the Browse For Folder window appears.



- a. (Optional) Select a different installation folder.
 - b. (Optional) Click **Make New Folder**.
 - c. Click **OK**.
4. Compare **Space required** with **Space available** in the Choose Install Location page. Ensure that there is enough room on your hard drive for the applications that you are installing.

-
5. Click **Next** and the Choose Components page appears.



6. (Optional) Click  to see all of the applications that you are installing and to deselect some, but not all, of these components:
- SAS Content Categorization Server
 - SAS Content Categorization Java API
 - SAS Content Categorization Python API
 - SAS Document Conversion Server and Java API

-
7. Click **Next** and one of the following SAS Content Categorization Single User Servers Setup windows might appear. If none of these windows appears, see Step 8. on page 20.

- If you selected SAS Content Categorization Java API and you do not have version 1.5 or later loaded, the SAS Content Categorization Single User Servers Setup window appears. Note the URL and load Java after the installation of SAS Content Categorization Single User Servers is complete:



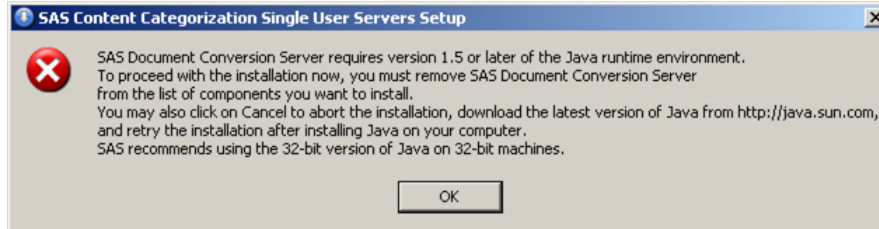
- If you selected SAS Content Categorization Python API and you do not have Python version 2.3 or higher loaded, the SAS Content Categorization Single User Servers Setup window appears. Note the URL and load the recommended Python version and type before you install SAS Content Categorization Python API.

Hint: You can deselect this operation in order to continue the Install operation and install SAS Content Categorization Python API later.

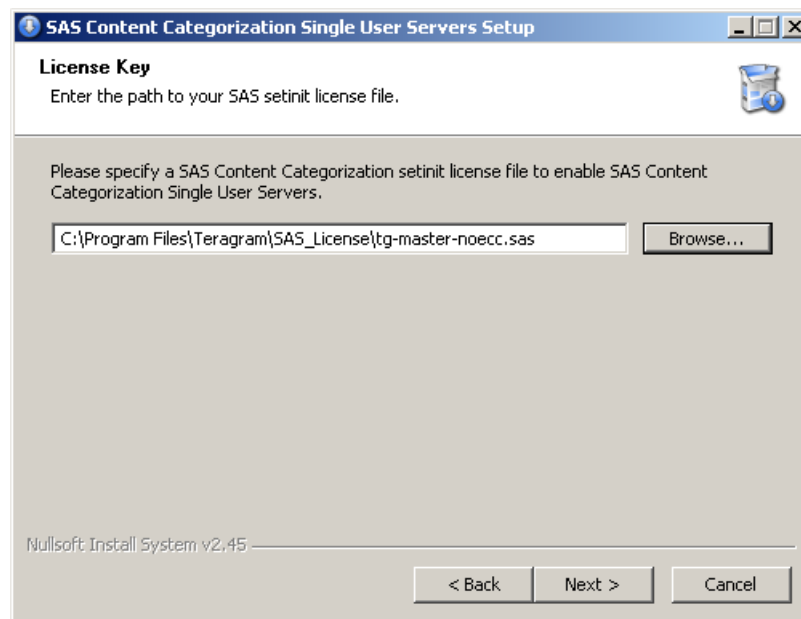


- (If you select SAS Document Conversion Server and Java API and you do not have version 1.5 of the Java run-time environment loaded, deselect this component.) If, instead you select this

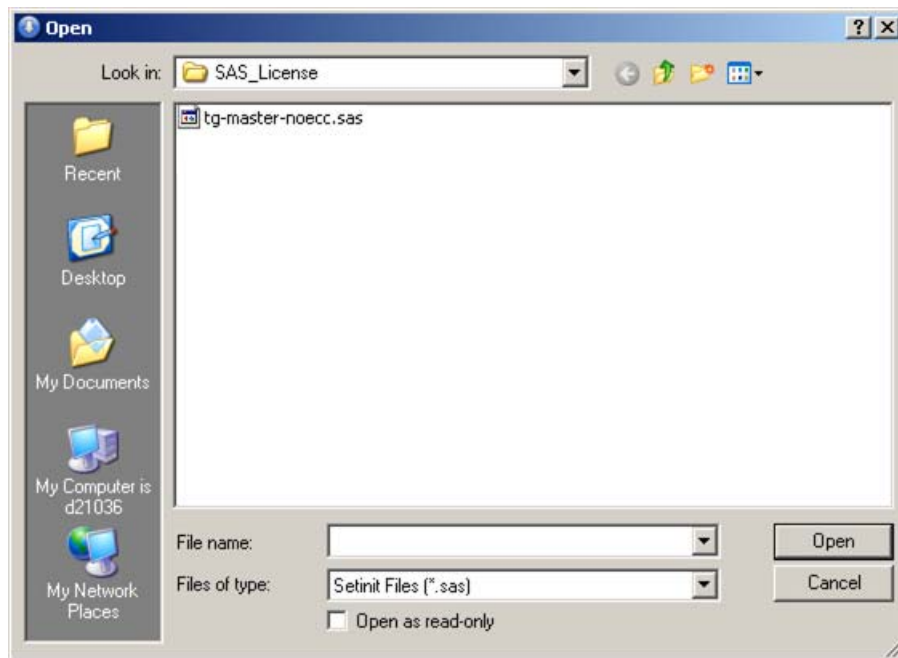
component, the Setup window appears. Note the URL and load Java before you try to install SAS Document Conversion Server:



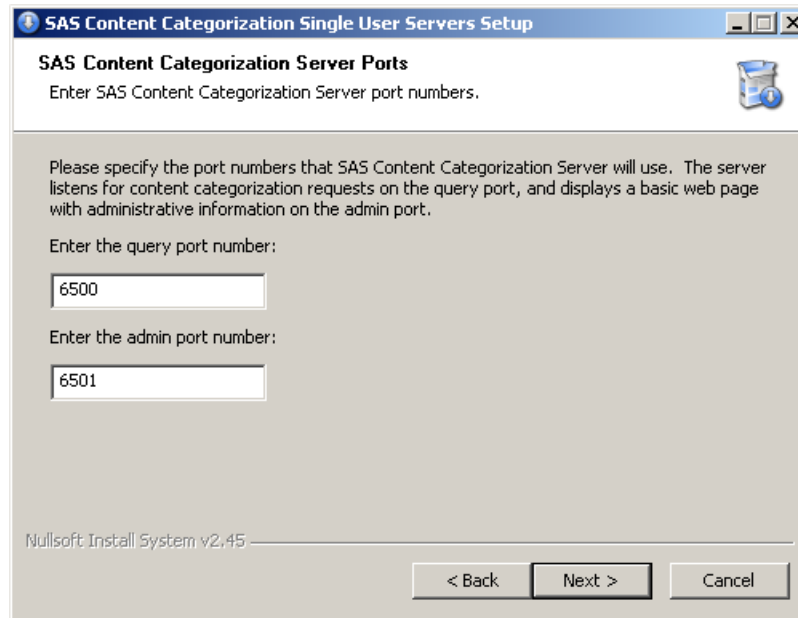
8. Click **OK** in any of the SAS Content Categorization Single User Servers Setup windows that appear. See the License Key page that appears:



-
9. Enter the path to the license key or click **Browse**. If you click **Browse**, see Step 3. on page 17. Specify this path on a local drive, not a network drive.



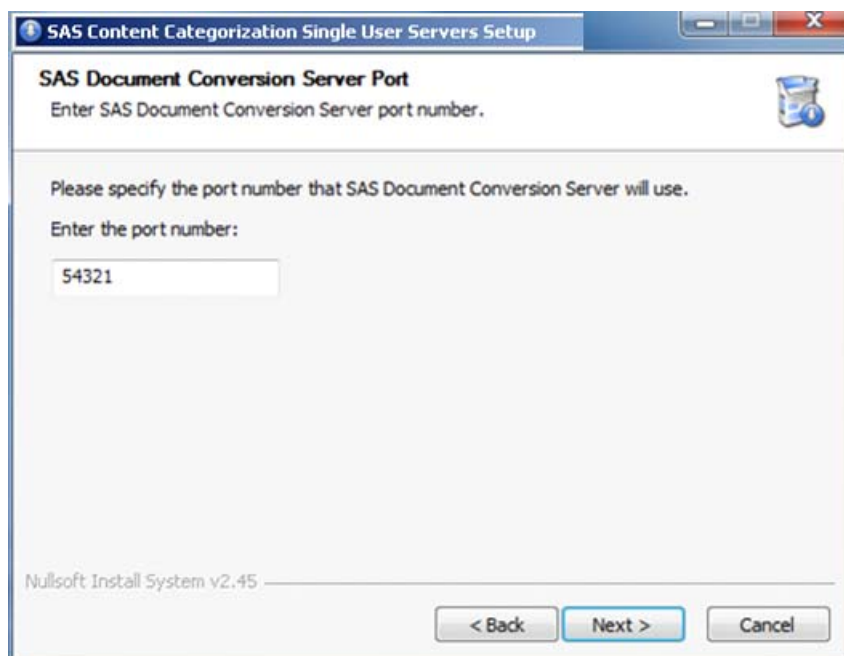
-
10. Click **Next** and the SAS Content Categorization Server Ports page appears.



Note: This page is only for SAS Content Categorization Server.

11. (Optional) Use the **Enter the query port number** field to change the port number that appears by default.
12. (Optional) Use the **Enter the admin port number** field to change the port number that appears by default.

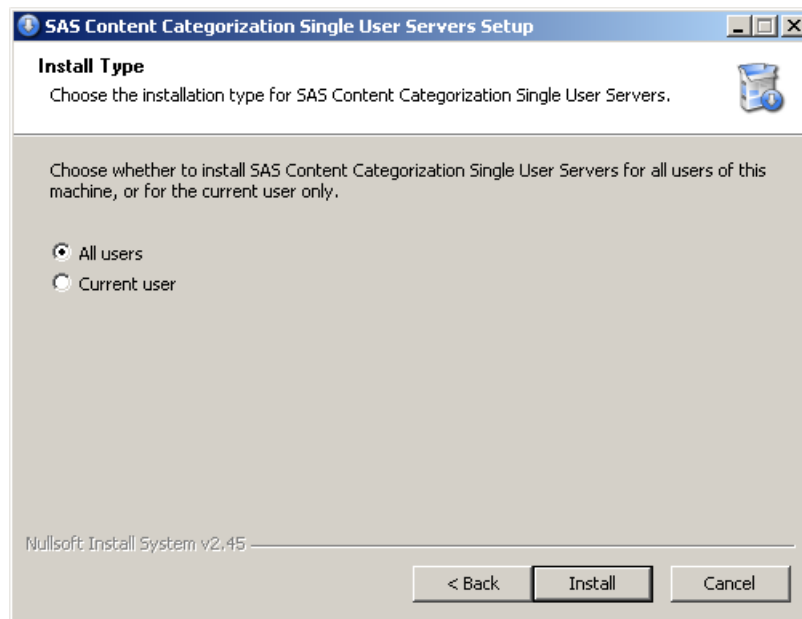
-
13. Click **Next** and the SAS Document Conversion Server Port page appears.



The screenshot shows a Windows-style dialog box titled "SAS Content Categorization Single User Servers Setup". The main heading is "SAS Document Conversion Server Port". Below the heading, it says "Enter SAS Document Conversion Server port number." and "Please specify the port number that SAS Document Conversion Server will use." There is a text input field labeled "Enter the port number:" containing the value "54321". At the bottom, there are three buttons: "< Back", "Next >", and "Cancel". The "Next >" button is highlighted. The footer of the window says "Nullsoft Install System v2.45".

14. (Optional) Use the **Enter the port number** field to change the port number that appears by default.

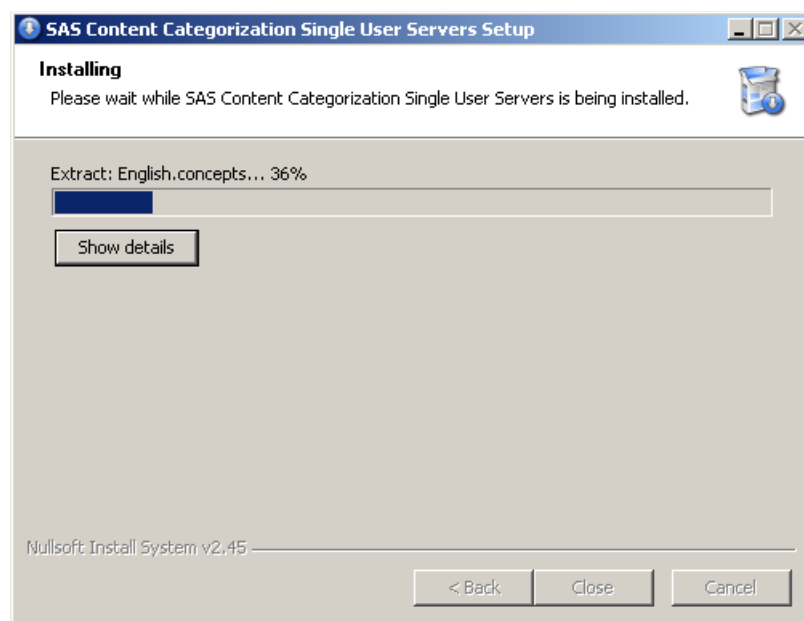
15. Click **Next**. The Install Type page appears:



16. Install the servers for either:

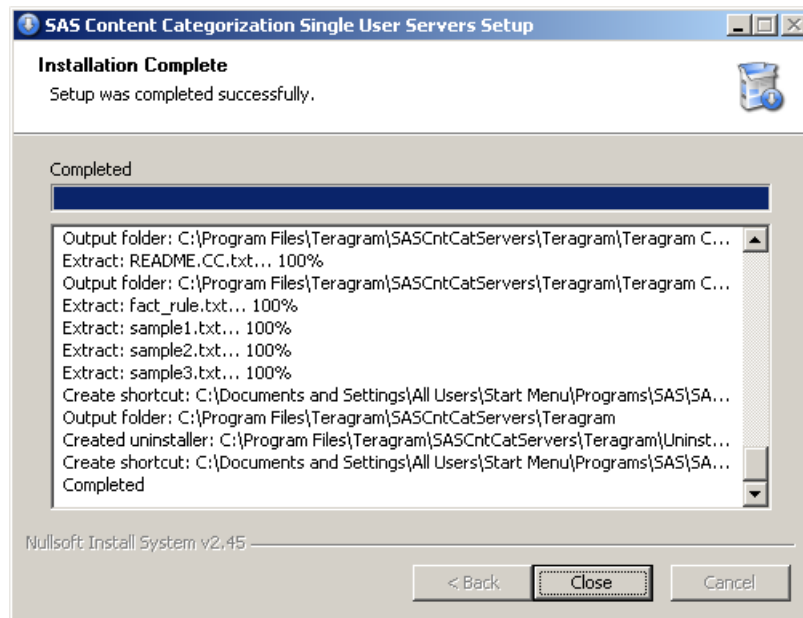
- **All users:** This operation enables all of the users on this machine to access the servers.
- **Current user:** This operation enables only the person who performs the Install operation to access the downloaded servers.

-
17. Click **Install** and the Installing page appears where you can see the installation progress.



18. (Optional) Click **Show details** and see Step 19. below for an example of the pane that appears.

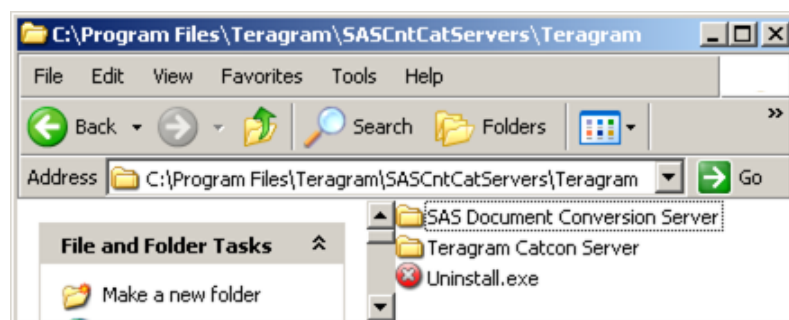
19. Click **Close** in the Installation Complete page that appears.



2.2.1.C The Folders That Appear after Installation

See the following example of the folders that might appear after installation:

Display 2-1 Installed Folders



SAS Document Conversion Server

find the empty `client api` folder and a `README` file.

Teragram Catcon Server

find a number of files and folders, including the SAS Content Categorization Server configuration file in the `conf` directory.

`uninstall.exe`

uninstall SAS Content Categorization Servers.

2.2.2 Install on UNIX

2.2.2.A Install on UNIX

SAS Content Categorization Single User Servers is distributed on UNIX systems as a `tar.gz` file:

```
SAS_Content_Categorization_Server--<arch>--  
    <date>.tar.gz
```

<arch>

is a string representing the architecture. For example, see `linux64`.

<date>

is the release date of the package.

To install the software, use the following UNIX commands:

```
gzip -d sas_cc_server_<arch>.tar.gz
tar -xvpf sas_cc_server_<arch>.tar
```

The switches on the `tar` command are used to extract the contents from the specified `tar` file and to preserve the file and directory permissions of the contents.

Note: The actual name of your `tar` file might vary from that shown in the example above.

Additional information about using the `gzip` and `tar` commands is available in the UNIX `man` pages.

2.2.2.B See the Directory Structure

When these files are extracted, the following directory structure is created. See the notes on what is in each directory in *italic font*.

```
+ sas_cc_servers_<arch>
|-- cc_server
+
|-- bin
+
  |-- <arch> (SAS Content Categorization Server
              executable)
  |-- client_api (README file for SAS Content
                  Categorization client APIs)
+
  |-- java (Java client API)
+
  |-- doc (Java client API technical documentation)
  |-- python (Python client API)
+
  |-- doc (Python client API technical documentation)
  |-- test (Sample documents for testing the client
            APIs)
|-- conf (server.config file)
|-- descriptors (Descriptor files to load the included
                 sample .mco and .concepts files)
|-- doc (SAS Content Categorization Single User
         Servers: Administrator's Guide)
```

```
|-- models (Sample .mco and .concepts files)
|-- doc_conversion_server (Java JAR files for SAS
    Document Conversion Server and client)
+
|-- doc (README file for SAS Document Conversion Server
    and client API)
+
|-- javadoc (Java client API technical documentation)
|-- open-source (obligatory legal disclaimers for use
    of the open source library that SAS Document
    Conversion Server 12.1 )uses)
```

2.3 Starting and Stopping a Server on a Windows Machine

After you install SAS Content Categorization Single User Servers, SAS Content Categorization Server automatically starts. The following steps are for SAS Content Categorization Server, but you can modify these steps for SAS Document Conversion Server.

To stop SAS Content Categorization Server, complete the following steps:

1. Go to **Start --> Settings --> Control Panel --> Administrative Tools --> Services**.
2. Select the SAS Content Categorization Server.
3. Right-click and select **Stop** from the drop-down menu that appears.

To restart SAS Content Categorization Server, complete the following steps:

1. Complete Step 1. and Step 2. above.
2. Right-click and select **Start** from the drop-down menu that appears.

Note: You can access the SAS Content Categorization Server Administration Web Page only from the **Start** menu when the server is running. For more information, see Chapter 5: *SAS Content Categorization Server Web Administration*.

2.4 Uninstall on Windows

Administrators, or users with administrative permissions, are the only users that can uninstall SAS Content Categorization Single User Servers.

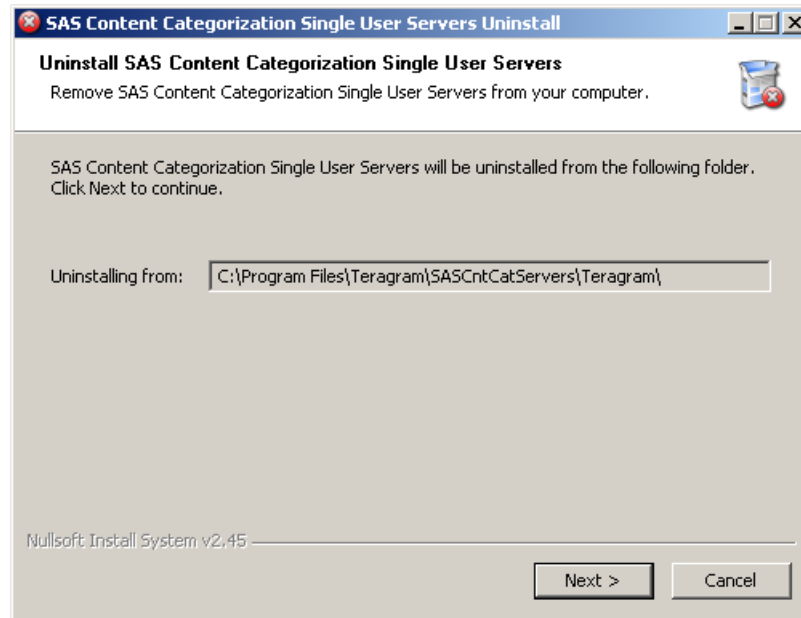
Notes: Before you perform the Uninstall operation, stop SAS Content Categorization Server and SAS Document Conversion Server. For more information, see Section 2.3 *Starting and Stopping a Server on a Windows Machine* on page 29.

If you install SAS Content Categorization Single User Servers as an administrator, you are the only person who can uninstall this software.

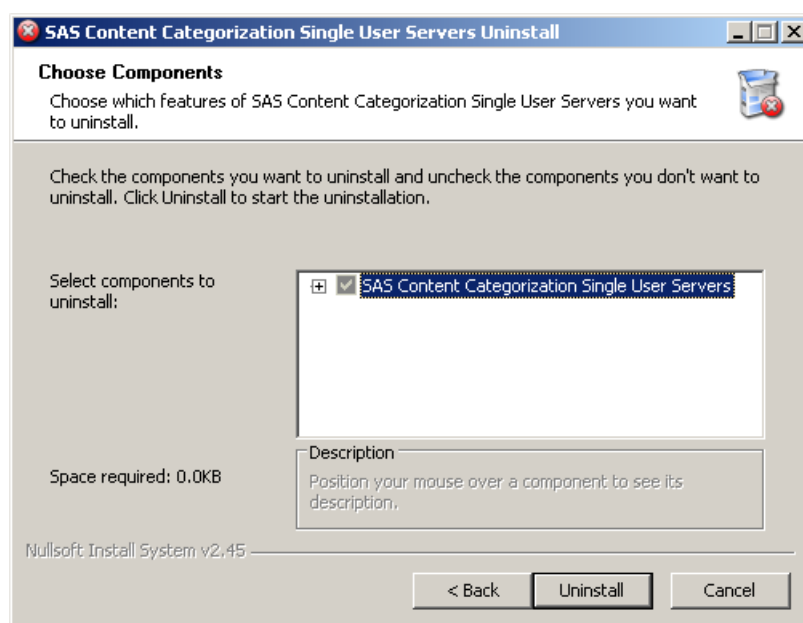
To uninstall SAS Content Categorization Single User Servers, complete these steps:

1. Go to **Start --> Programs --> SAS --> SAS Content Categorization Single User Servers --> Uninstall SAS Content**

Categorization Single User Servers. The Uninstall SAS Content Categorization Single User Servers page appears.

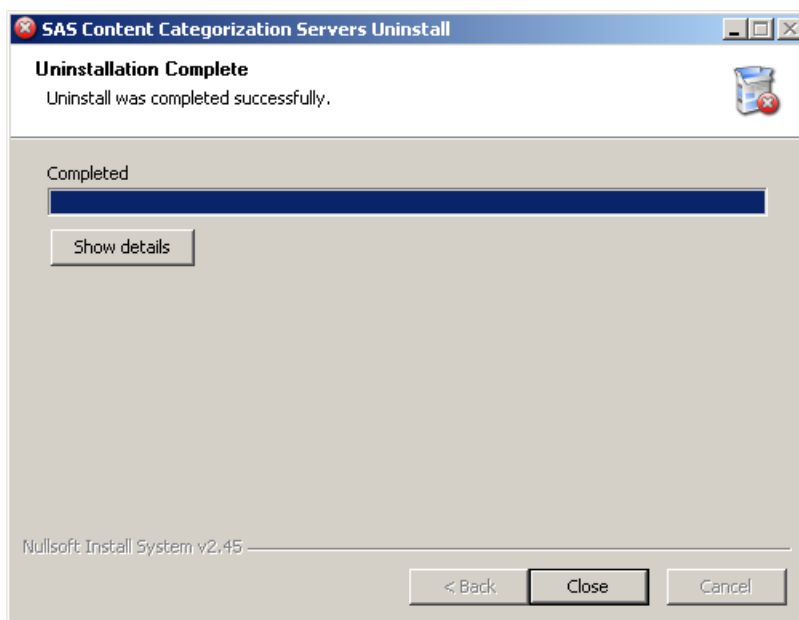


-
2. Click **Next**. The Choose Components page appears:



3. (Optional) By default, all of the components that you installed are checked. Uncheck the components that you do not want to uninstall.

-
4. Click **Uninstall**. The Uninstall page appears after the server stops running.



5. (Optional) Click **Show Details**. See Step 19. on page 26 for an example of the pane that appears.
6. Click **Close**.

2.5 Access the Servers

To use the Start menu to locate the SAS Content Categorization Single User Servers program, complete these steps:

-
1. Go to **Start --> Programs --> SAS --> SAS Content Categorization Single User Servers**:



2. Select any of the following:
 - **Documentation:** See the documentation that is available for SAS Content Categorization Single User Servers.
 - **Configure SAS Content Categorization Server:** Access the configuration file, if you selected this component.
 - **SAS Content Categorization Server Administration Web Page:** Check SAS Content Categorization Server operations using this Web page.
 - **Uninstall SAS Content Categorization Single User Servers:** Remove the selected components from your machine.

Hint: If you did not install SAS Content Categorization Server, only the **Documentation** and **Uninstall SAS Content Categorization Servers** selections are available.

2.6 Processing Documents

You can upload a project that uses the new synonym list feature in SAS Content Categorization Studio. In this case, make sure that the documents that you send to SAS Content Categorization Single User Servers do not exceed 1 MB in size.

Part 2: SAS Content Categorization Server

- Chapter 3: *Configuring and Running SAS Content Categorization Server on page 37*
- Chapter 4: *SAS Content Categorization Server Web Administration on page 57*

Chapter: 3

Configuring and Running SAS Content Categorization Server

- *Configuration Overview*
- *SAS Content Categorization Server Configuration File*
- *Specifying Project Users and Creators*
- *Add a Project*
- *Using the Models Directory*
- *Using the Descriptors Directory*
- *Using cat_log and concept_log Files*
- *Specifying Multiple Project Files*
- *Sending Documents to the Server*
- *Run SAS Content Categorization Server*
- *Optimize Performance on a Client Windows Machine*

3.1 Configuration Overview

Configure SAS Content Categorization Server by specifying directives in the configuration file (`server.config`). The configuration file is automatically created when you install SAS Content Categorization Server and the file remains after the uninstall operation.

These directives enable you to specify creators who are administrators who can upload projects and other users for this application. You also choose the types of connections for the server, directories, and other settings. The configuration file is automatically created when you install SAS Content Categorization Server and remains after the uninstall operation.

Note: If you are installing SAS Content Categorization Server for the first time, the configuration file is the default file. For more information, see Section 3.2 *SAS Content Categorization Server Configuration File* on page 39. If you have previously installed SAS Content Categorization Server, the original configuration file is preserved. For more information about how to edit this file, see Appendix A: *Troubleshooting*.

The SAS Content Categorization Server configuration file is a text file that contains key - value pair assignments. See the following form, where each pair appears on a single line. Any blank lines, as well as any comment lines that are preceded by the comment character (#), are ignored.

key=value

The `descriptor_dir` directive specifies the directory that contains descriptor files. For example, see the `descriptors` directory. The descriptor files determine the binary models that are loaded and the symbolic names for these models. When models are uploaded to SAS Content Categorization Studio, these files are automatically created.

See the following descriptor file examples:

Example 3-1: ITPC.desc

```
type=mcats
path=models\English.mco
name=ITPC
```

Example 3-2: Entities.desc

```
type=concepts
path=models\English.concepts
name=Entities
```

Note: The UNIX path contains a forward slash (/) instead of a backslash.

3.2 SAS Content Categorization Server Configuration File

3.2.1 Windows Configuration File

During installation, the configuration file is automatically created in the following location:

```
<INSTALL_DIR>\Teragram Catcon Server\conf
```

To see this file, select **Start --> Programs --> SAS --> SAS Content Categorization Single User Servers --> Configure SAS Content Categorization Server**. The configuration file that appears is similar to the example shown below:

Example 3-3: Sample Configuration File for Windows

```
basedir=C:\Program Files\Teragram\SASCntCatServers\
      Teragram\Teragram Catcon Server\
backupdir=backup
setinit=C:\Program Files\Teragram\SAS_License\tg-
      master-noecc.sas
descriptor_dir=descriptors
create_dir=models
query_port=6500
admin_port=6501
skt_queue_size=10
nb_threads=4
persistent_connection=0
timeout=60000000
max_iterations_to_reinitialize=5000
```

3.2.2 UNIX Configuration File

On a UNIX system, the configuration file is located in the `conf` subdirectory. For example:

```
/opt/sas_catcon_server_linux64/conf/server.config
```

3.2.3 The Directives

Use the directives to modify the SAS Content Categorization Server configuration file. Directives such as `descriptor_dir` are treated as absolute paths. In other words, these paths are not relative to `basedir`. This is true if `basedir` is specified and these paths begin with `[A-Z, a-z]:\` on Windows or on UNIX. See the following example:

```
backupdir=c:\backups
```

The directives for the SAS Content Categorization Server configuration file are described in the table below:

Table 3-1: Configurable Directives

Directive	Description
<code>basedir</code>	Specifies the path to the project binaries, backup directory, and so on.
<code>backupdir</code>	Specifies the directory where the backup binaries are stored. When a categorization (<code>.mco</code>) or concepts (<code>.concepts</code>) binary file is opened by SAS Content Categorization Server at start-up and this directive is specified, a backup copy of the binary file is created. The backup copy is written to the specified directory. If a binary file cannot be subsequently opened, SAS Content Categorization Server attempts to use the backup version of the binary file. This directive enables SAS Content Categorization Server to keep running even if a binary file cannot be loaded.
<code>setinit</code>	Specifies the path to the SAS license file for SAS Content Categorization Server. This SAS installation data file (SID) is in the Software Order E-mail (SOE) that you received. For more information, see Section 2.1.3 <i>Using the SAS License File</i> on page 13.
<code>descriptor_dir</code>	Contains descriptor files that contain information about the projects that are loaded on SAS Content Categorization Server. Hint: After you uninstall SAS Content Categorization Server, this directory and its files remain. This statement is true if you have uploaded one or more models to the server.

Table 3-1: Configurable Directives (Continued)

Directive	Description
create_dir	Tells SAS Content Categorization Server where to store the binary files for the projects that are added to this server. Without this directive, the creator cannot upload new files. Hint: After you uninstall SAS Content Categorization Server, this directory and its files remain. This statement is true if you have uploaded one or more models to the server.
query_port	Specifies the number of the TCP port where the categorization and concept extraction services are available. The clients connect to this port on the server host. This port number is used in the Port field of the Upload concepts to SAS Content Categorization Studio window. For more information, see Section 3.4 <i>Add a Project</i> on page 45.
admin_port	Specifies the number that corresponds to the TCP port where the server's Web-based administrative interface is available.
skt_queue_size	Specifies the number of simultaneous pending client connections that the server accepts, without dropping the connection. If all of the server threads are busy serving clients, this attribute specifies the maximum number of additional clients. This number of clients can connect to the server and wait for a thread to become available.
nb_threads	Specifies the number of parallel service threads to run. SAS Content Categorization Server is able to handle the specified number of clients, simultaneously.
persistent_connection	Specifies whether the server tries to maintain a continuous socket connection with the client, or not. The default value is zero (0). If this setting is set to one (1), persistent connections are enabled if the client also enables these connections.
timeout	Specifies the length of time (in microseconds) that the server waits. If no activity occurs during this period, the server forcibly drops the connection.
max_iterations_to_reinitialize	Tells the server to clear out its memory after the specified number of documents is reached.
max_doc_size	Specifies the largest size (in bytes) of documents that can be processed. Larger texts are truncated.

Table 3-1: Configurable Directives (Continued)

Directive	Description
xml_weight_file	<p>Specifies the weights for structured-text fields that match MCAT rules. When rule terms match within these fields, the relevancy score for these terms is multiplied by the field weight. The syntax for the <code>xml_weight_file</code> is <code>field: weight</code> for each line in the file. You could specify the following:</p> <pre>title:3 body:1.5</pre> <p>In this example, if a match is located in the <code>body</code> field of an XML document, the match counts 1.5 times toward the relevancy score. However, a match for the <code>title</code> field is multiplied by three.</p>
user	<p>Specifies a user name and password. Users specify these entries to upload the binary files directly from SAS Content Categorization Studio to SAS Content Categorization Server. These users can upload only the new versions of projects that already exist on the server. For more information, see Section 3.3 <i>Specifying Project Users and Creators</i> on page 44.</p> <p>Tip: This directive can be specified multiple times.</p>
creator	<p>Specifies a user name and password necessary to upload new projects and refresh existing projects. For more information, see Section 3.3 <i>Specifying Project Users and Creators</i> on page 44.</p>
io_log	<p>Generates a detailed input and output log file while performing the categorization operation. This log includes timestamps. The value is the name of the file where the logging information is written.</p> <p>Tip: This directive only applies to categories.</p>
cat_log	<p>Generates a log of all of the category matches that are returned for the documents sent to the server. One entry is specified for each document that matches one or more categories in a category project. The value is the base for the category log file. For more information, see Section 3.7 <i>Using cat_log and concept_log Files</i> on page 50.</p>
cat_log_max_entries	<p>Specifies the maximum number of entries allowed in each <code>cat_log</code> file. For more information, see Section 3.7 <i>Using cat_log and concept_log Files</i> on page 50.</p>

Table 3-1: Configurable Directives (Continued)

Directive	Description
num_cat_logs	Specifies the maximum number of category log files to create. For more information, see Section 3.7 <i>Using cat_log and concept_log Files</i> on page 50.
do_cat_log_timing	Provides additional timing information in all of the category log files. No value is required.
concept_log	Generates a log of all of the concept matches that are returned for the documents sent to SAS Content Categorization Server. One entry is defined for each document that matches one or more concepts in a concepts project. The value is the base for the filename. For more information, see Section 3.7 <i>Using cat_log and concept_log Files</i> on page 50.
concept_log_max_entries	Specifies the maximum number of entries allowed in each concept_log file. For more information, see Section 3.7 <i>Using cat_log and concept_log Files</i> on page 50.
num_concept_logs	Specifies the maximum number of concept log files to create. For more information, see Section 3.7 <i>Using cat_log and concept_log Files</i> on page 50.
do_concept_log_timing	Obtains additional timing information in the concept log file. It is not necessary to specify any value for this attribute.
protocol_version	Enables SAS Content Categorization Server to emulate older versions of the client/server protocol that this application uses.

Note: After modifying and saving a configuration file, restart SAS Content Categorization Server in order to make the changes take effect. For more information, see Section 2.3 *Starting and Stopping the Servers on a Windows Machine* on page 27.

The directives in the following table continue to be supported for backwards compatibility purposes. Use the `descriptor_dir` directive when possible.

Table 3-2: Deprecated Directives

Directive	Description
<code>mcat</code>	Specifies either a relative, or an absolute, path to a categorization binary file (<code>.mco</code>). The symbolic name for the categories project is specified after a colon (<code>:</code>).
<code>concepts</code>	Specifies either a relative or an absolute path to a concepts binary file (<code>.concepts</code>). The symbolic name for the concepts project is specified after a colon (<code>:</code>).
<code>stat_cat</code>	Specifies either a relative or an absolute path to a statistical categorizer binary file (<code>.st.cat</code>). The symbolic name is specified after the colon (<code>:</code>).
Tips: These directives can be specified multiple times. If <code>basedir</code> is specified, these paths are relative to this directory. If this directory is not specified, these paths are absolute.	

3.3 Specifying Project Users and Creators

In order to upload a project file from SAS Content Categorization Studio to SAS Content Categorization Server, specify a `creator` in the server configuration file. You can also add users who have permissions to refresh existing project files to the server.

User

refresh a binary file for a project that is already uploaded to SAS Content Categorization Server. The user can reload these project files from SAS Content Categorization Studio to SAS Content Categorization Server.

Creator

is an administrator who can upload a binary file for a new project to SAS Content Categorization Server. As an administrator, you create new projects on this server.

Using both directives, or only the `creator` directive, specify your name and password. These specifications apply to any of the Upload operations in SAS Content Categorization Studio.

An example of the directives for the `user` and `creator` that you can add to the configuration file are shown below:

Example 3-4: User and Creator Directives

```
user=user1:pw1  
creator=creator1:pw3
```

All duplicate usernames are ignored after the first instance. See the following example:

```
creator=Joe:Joespassword  
user=Joe:Joespassword
```

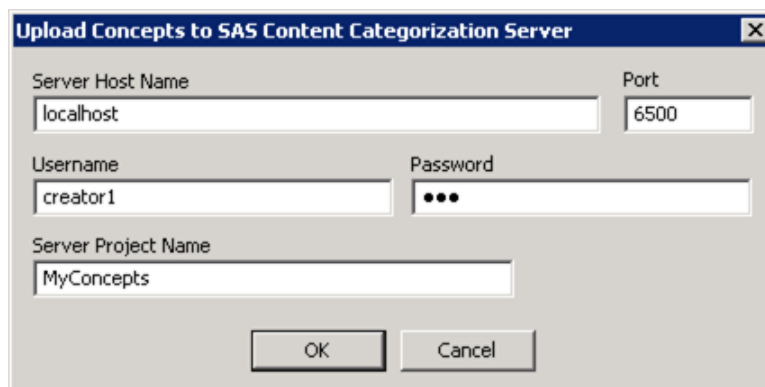
In this example, `Joe` has creator permissions. If you specify `-verbose`, the server emits a warning for duplicate or conflicting entries.

3.4 Add a Project

Administrators who have creator privileges can upload projects built in SAS Content Categorization Studio to SAS Content Categorization Server. When a creator uploads a project, files for the project is automatically added to the data and descriptor directories. These folders are referenced by the configuration file for SAS Content Categorization Server. For this reason, it is not necessary to add new projects directly to the configuration file.

To add a `.concepts` project developed in SAS Content Categorization Studio to SAS Content Categorization Server, complete these steps: (Make appropriate changes for `.mco` files.

-
1. In SAS Content Categorization Studio, select **Build --> Upload Concepts**. The Upload Concepts to SAS Content Categorization Studio window appears:



The screenshot shows a dialog box titled "Upload Concepts to SAS Content Categorization Server". It contains the following fields and values:

- Server Host Name: localhost
- Port: 6500
- Username: creator1
- Password: (masked with three dots)
- Server Project Name: MyConcepts

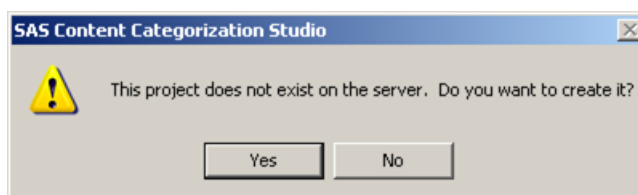
At the bottom of the dialog are "OK" and "Cancel" buttons.

By default, the **Server Host Name** and **Port** field values are automatically entered.

Note: The value for the **Port** field should be the value specified in the `query_port` directive of the `server.config` file, not the `admin_port` field. For more information, see Example 3-3 on page 39.

2. Enter the creator name that is specified in the configuration file into the **Username** field. For example, type `creator1`.
3. Enter your creator password that is specified in the configuration file into the **Password** field. For example, type `pw3`.
4. Enter the symbolic name for the file that you choose to upload into the **Server Project Name** field. For example, type `MyConcepts`.

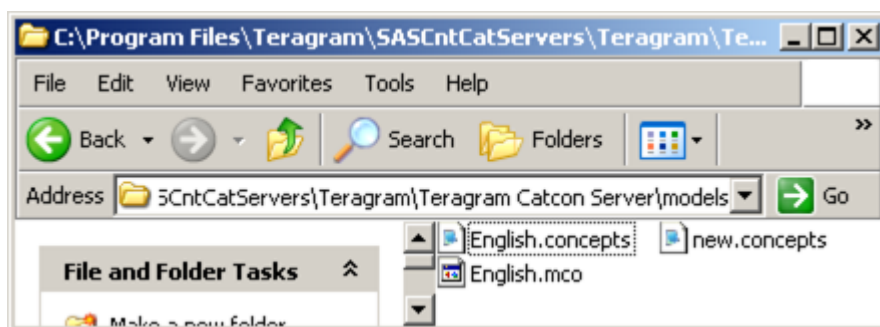
-
5. Click **OK**. A SAS Content Categorization Studio status window appears.



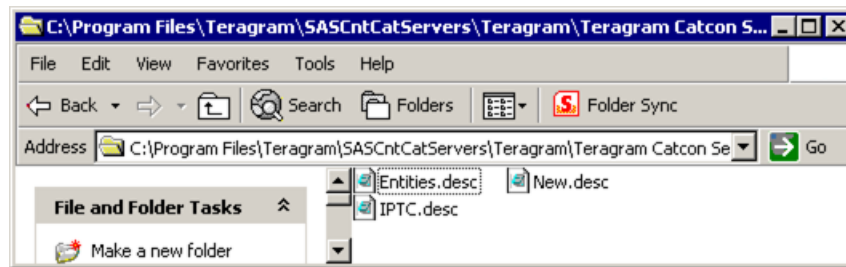
6. Click **Yes**. A second SAS Content Categorization Studio status window appears.



7. Click **OK** and the project is listed in the `models` folder. For more information, see Section 3: *Using the Models Directory*.



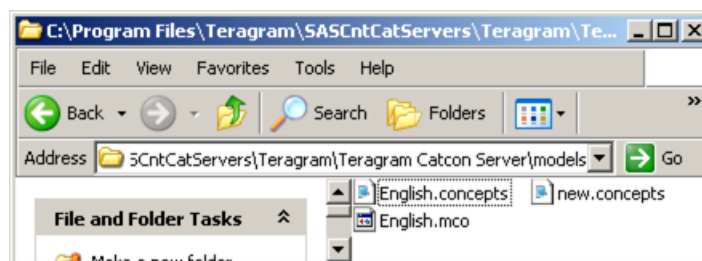
-
8. The project is also listed in the `descriptors` folder. For more information, see Section 3: *Using the Descriptors Directory*.



3.5 Using the Models Directory

The `models` directory displays each of the projects that are uploaded to SAS Content Categorization Server. These projects include any sample projects that are shipped with the application. For example, the names of the shipped project files might be `English.concepts` and `English.mco`. You also might have uploaded a project file such as `new.concepts` with this installation or with an earlier installation. In either case, this file also appears in the folder.

Display 3-1 The Models Directory

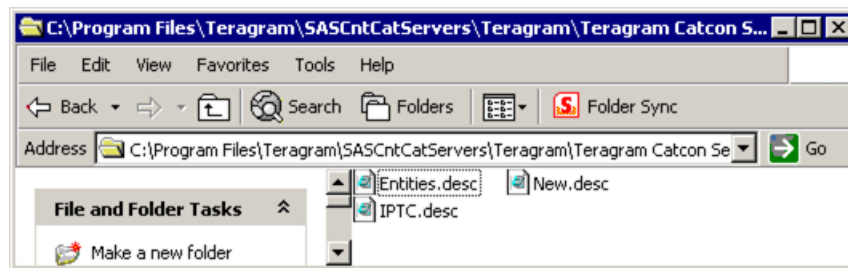


3.6 Using the Descriptors Directory

The descriptor directory (`descriptors`) contains information about each project that is uploaded to SAS Content Categorization Server. This information is in the form of binary files called descriptor files that follow the `<project name>.desc` naming convention. These files are automatically created for each project that is uploaded. Without the `descriptor` directory, no project files can be uploaded.

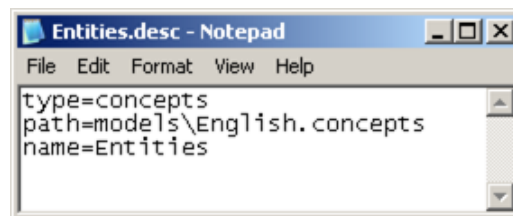
A plain text `.desc` file is created each time a creator uploads a project, unless a project with a duplicate name is uploaded. In this case, the existing file is overwritten. See the following example:

Display 3-2 The Descriptors Directory



Double-click the `.desc` file to access and read the file contents. The format for these files is displayed in the `Entities.desc` example that is shown below:

Display 3-3 Sample Descriptor File



Explanations for the lines in this file that are shown below:

`type=[mcat | concepts]`

specifies the type of project that is uploaded. You can upload a categories (.mco) or a concepts (concepts) file.

path=<file_location>

specifies the data path relative to the base, or installation, directory. If basedir is not specified in the configuration file, this path is treated as an absolute path.

name=<symbolic_name>

specifies the name that you assign to the project.

Notes: The file path is relative only to basedir. This path is not relative to basedir combined with create_dir. The deprecated mcat and stat_cat directives can co-exist with the descriptor_dir directive.

3.7 Using cat_log and concept_log Files

There are several ways that you can use the cat_log files. These examples also apply to concept_log files. The word *entry* is defined here in the context of the SAS Content Categorization Server configuration file. *Entry* represents a document that matches one or more categories or concepts within the project.

Note: If you want to use the SAS Content Categorization Server Administration Web Page, specify the appropriate log files in your server.config file. For more information, see Table 3-1 on page 40.

For example, if an input text matches one category in each of two projects, two entries are created. If however, another text matches five categories in one project and two in another project, two entries are also created. (For information about loading multiple project files into your SAS Content Categorization Server, see Section 3.8 *Specifying Multiple Project Files* on page 51.)

Your SAS Content Categorization Server configuration file might contain the following line:

```
cat_log=cat.log
```

In this example, the category log file is named `cat.log.0`. In this case, because there is no specification for `cat_log_max_entries`, the number of entries continue to grow. This growth continues until the log file is deleted or until there is no more disk space. However, you can specify a maximum number of entries for this file using the following example:

```
cat_log=cat.log
cat_log_max_entries=10000
```

In this example, the `cat.log.0` file is regenerated whenever more than 10,000 documents match at least one category, in at least one project, and during one session. When the `cat.log.0` file is regenerated, all existing data in the file is lost. For this reason, you can also configure SAS Content Categorization Server to create more than one log file. See the following example:

```
cat_log=cat.log
cat_log_max_entries=10000
num_cat_logs=10
```

To begin the regeneration process, SAS Content Categorization Server creates a file named `cat.log.0`. The server might attempt to exceed the specified number of entries. In this example 10,000 entries are specified. In this case, `cat.log.0` is copied to `cat.log.1`, and `cat.log.0` is regenerated to include the excess entries. This process can continue until the limit of `cat.log.9` is reached. In this case, excess data is copied and the first log file is destroyed when the maximum number of entries is reached.

3.8 Specifying Multiple Project Files

You can choose to load as many project files to SAS Content Categorization Server as your system can hold in memory. If you update your projects in SAS Content Categorization Studio, you can reload the updated projects using the same symbolic name specified for the original project.

The first project loaded on the server, or the first descriptor file that is read, for each binary file type is the default project. These project names are also used by the APIs for SAS Content Categorization Server.

3.9 Sending Documents to the Server

In order to send documents to SAS Content Categorization Server, you use the Java programs described in the README file. This file is located in the `client_api` directory that use the Java API. See the path below:

`Teragram Catcon Server\client_api\java`

3.10 Run SAS Content Categorization Server

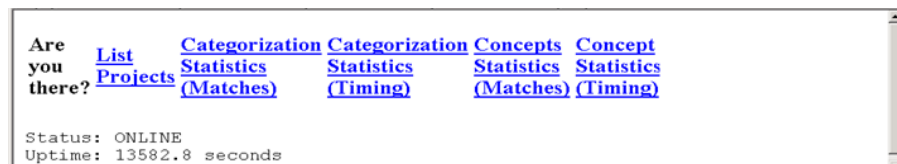
3.10.1 Verify That the Server Is Running

You can check to see whether SAS Content Categorization Server is running using the Administration Web Page or by checking the Services running on your machine.

To check whether the SAS Content Categorization Server service is running using the Web page, complete these steps:

1. Go to **Start --> Programs --> SAS --> SAS Content Categorization Single User Servers --> SAS Content Categorization Server Administration Web Page.**

Display 3-4 The Administrative Web Page



2. Click **Are you there?**
3. See the Status and Uptime. When the server is running the status is ONLINE and Uptime displays a number of seconds.

To stop or restart SAS Content Categorization Server, use the directions in Section 2.3 *Starting and Stopping the Servers on a Windows Machine* on page 27.

3.10.2 Run the Server on UNIX

To run SAS Content Categorization Server on a supported UNIX system, go to the installation root directory and enter the following command from the UNIX shell:

```
# ./bin/<arch>/-catcon_server.exe
```

For example:

```
# ./bin/linux64/_catcon_server
```

In this command line, `-server configfile` specifies the name and full path to the configuration file. The server program runs in the foreground. This means that it does not fork and writes its logging output to the terminal that initiated the program (`stdout`).

Note: Use the `-verbose` switch for debugging purposes.

3.11 Optimize Performance on a Client Windows Machine

3.11.1 Overview of Performance Optimization

The settings that are specified in the following sections should be applied if the client program that connects to SAS Content Categorization Studio is running on Windows. Otherwise, unexpected behaviors might occur when you process large amounts of documents.

3.11.2 Before and After You Optimize Performance

Before you use the following sections to optimize the performance of SAS Content Categorization Server, run the registry editor.

To run the registry editor, complete these steps:

1. Select **Start --> Run**.

-
2. Type `regedit` into the **Open** field of the Run window that appears.
 3. Click **OK**.
 4. After you use both Section 3.11.3 *Adjust the TCP Time Wait State* below and Section 3.11.4 *Reset Ephemeral Ports* on page 54, reboot your machine.

3.11.3 Adjust the TCP Time Wait State

Choose to lower the setting for the timed wait state in order to avoid depleting available ports on your servers. SAS recommends that you consider setting this selection in your system registry to 15 seconds.

To reset the `TcpTimedWaitDelay` setting, complete these steps:

1. Go to the registry subkey:

```
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\
    Tcpip\Parameters
```

2. Select **Edit --> New --> DWORD value**. By default, the new value is named `New Value #1`.
3. Rename the value by entering `TcpTimedWaitDelay`.
4. Double-click on the new `TcpTimedWaitDelay` value.
5. Select `Decimal` as the base, and enter 15 for the value data.

3.11.4 Reset Ephemeral Ports

Ephemeral ports are short-lived ports that are used to create connections to the client computers from the server and between COM server objects. By default, these ports range from 1024 to 5000. Connection difficulties can occur if you run short of ports.

This section explains how to reset the parameter that controls the maximum port number that is used when the SAS Content Categorization Server program requests an available user port from the system.

Use the following steps to reset the valid range for ephemeral ports:

1. Go to the registry subkey:

HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\
Tcpip\Parameters

2. Select **Edit --> New --> DWORD value**. By default, the new value is named `New Value #1`.
3. Rename the value by entering `MaxUserPort`.
4. Double-click on the new `MaxUserPort` value.
5. Select `Decimal` as the base, and enter `65534` for the value data.

Chapter: 4

SAS Content Categorization Server Web Administration

- *Overview of SAS Content Categorization Server Web Administration*
- *Access the Administration Web Interface*
- *Using the Administration Web Page*

4.1 Overview of SAS Content Categorization Server Web Administration

Use the SAS Content Categorization Server Administration Web Page while SAS Content Categorization Server is running, to see information about category matching and concept extraction. This interface can be used by either the server or creator administrators, or by a regular user.

To use the features of the Web Administration Page, specify the appropriate log files in your `server.config` file. For more information, see Table 3-1 on page 40.

The Administration Web Page enables you to perform the following tasks:

- Check to see that SAS Content Categorization Server is running.
- See a list of all of the loaded category and concept extraction projects.
- Use tables of statistics to analyze the categorization and concept extraction results.

Note: The statistics generated for category matches and concept extraction do not appear by default. To see these results specify the `cat_log` and `concept_log` lines

in the SAS Content Categorization Server configuration file.

For more information about the files that are loaded using the server configuration file, see Chapter 3: *Configuring and Running SAS Content Categorization Server*.

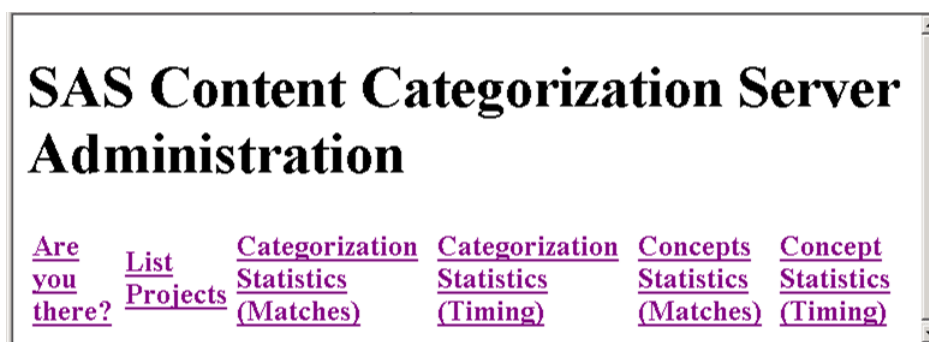
4.2 Access the Administration Web Interface

Use the links on the SAS Content Categorization Server Administration page to gain information about the documents that are processed by this server.

The SAS Content Categorization Server Administration page is displayed when you take the following step:

Go to **Start —> Programs —> SAS Content Categorization Single User Servers —> Start SAS Content Categorization Server Administration Web Page**.

The SAS Content Categorization Server - Administration page appears:



4.3 Using the Administration Web Page

4.3.1 Overview of the Administration Web Page

Use the Web Administration Page to see the operational data, in Web page format, as this information becomes available in SAS Content Categorization Server. Before completing the following steps, specify the `cat_log` or `concepts_log` directives in the server configuration file. For more information, see Section 3.7 *Using cat_log and concept_log Files* on page 50.

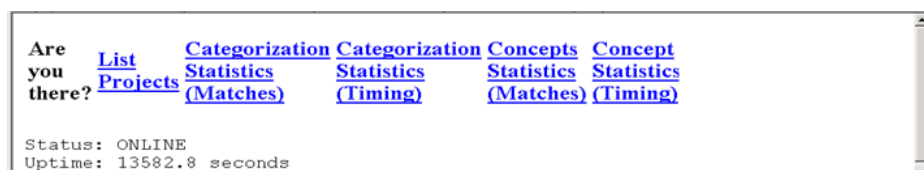
To run, access, and use the various management operations that are accessed through links on this page, use the following sections. Each operation has an assigned section in this chapter. The sections are ordered to match the links starting on the left side of the page and progressing to the right side of the page.

4.3.2 Use the Are you there? Page

The Are you there? page provides information about the status of the server and the length of time that the server was running.

To open and use the Are you there? page, complete these steps:

1. Click **Are you there?** and the Are you there? page appears.



2. See the following information about the server on this page:

Status

ONLINE, when running. Otherwise, there is no reply from the server. If the server is not started, see Section 2.3 *Starting and Stopping a Server on a Windows Machine* on page 29.

Uptime

displays the length of time that SAS Content Categorization Server has been running, or when the process was restarted. In the example above, the time is 2.2 seconds.

4.3.3 Use the SAS Content Categorization Server Projects List Page

The SAS Content Categorization Server Projects List page displays the categorization and concepts projects that are loaded on the server. This page also displays relevant information about each project such as the name and type of the project.

To use the SAS Content Categorization Server Projects List page, complete these steps:

1. Click **List projects** and the SAS Content Categorization Server Projects List page appears.

SAS Content Categorization Server Projects List

[Are you there?](#)[List Projects](#)[Categorization Statistics \(Matches\)](#)[Categorization Statistics \(Timing\)](#)[Concepts Statistics \(Matches\)](#)[Concept Statistics \(Timing\)](#)

Categorization Projects

	File Name	Symbolic Name	Type	Number of Rules	Default Category Bias	Default Relevancy Cutoff
0	models\English.mco	IPTC	Rule-based	1366	0	0.000000

Concepts Projects

	File Name	Symbolic Name	Number of Rules
--	-----------	---------------	-----------------

-
2. Use the tables in the SAS Content Categorization Server Projects List page to analyze information about the projects running on your server.

Table 4-1: Projects List Page Information






Heading	Description
Categorization Projects	
File Name	Specifies the name and location of the project file that was exported from SAS Content Categorization Studio. For example, the English.mco file located in the models folder.
Symbolic Name	Specifies the name of the project that you enter into the Server Project Name field of the Upload Categories to SAS Content Categorization Studio window.
Type	Specifies the rule type, which is either rule-based or statistical.
Number of Rules	Specifies the number of categories in this project.
Default Category Bias	Specifies the number that is set as the relevancy bias in SAS Content Categorization Studio for your categories. By default, this setting is set to 0.
Default Relevancy Cutoff	Specifies that any matching documents with a score below this number is not considered a match for a category. By default, this setting is set to 0.000000.
Note: You can change the Default Category Bias and the Default Relevancy Cutoff settings in SAS Content Categorization Studio Project Settings - Category window.	
Concepts Projects	
File Name	Specifies the name of the project file that was exported from SAS Content Categorization Studio. For example, this name could be English.concepts.
Symbolic Name	Specifies the name of the project that you enter into the Server Project Name field of the Upload Concepts to SAS Content Categorization Studio window.
Number of Rules	Specifies the number of concepts in this project.

4.3.4 Use the SAS Content Categorization Server Categorization Statistics (Matches) Page

The SAS Content Categorization Server Categorization Statistics (Matches) page lists the names of categories in the projects loaded onto the server. This page also displays information about the matches for these categories.

To use the SAS Content Categorization Server Categorization Statistics (Matches) page, complete these steps:

1. Click **Categorization Statistics (Matches)** and the SAS Content Categorization Server Categorization Statistics (Matches) page appears.

Are you there?	List Projects	Categorization Statistics (Matches)	Categorization Statistics (Timing)	Concepts Statistics (Matches)	Concept Statistics (Timing)
Total number of documents: 7					
Total number of categories with at least one match: 28					
Category Name	Percentage of Documents	Percentage (relative)			
Top/04000000 - Economy, Business and Finance/04003000 - Computing and Information Technology/04003004 - Semiconductors and Active Components	14.285714%				
Top/04000000 - Economy, Business and Finance	14.285714%				
Top/13000000 - Science and Technology	14.285714%				
Top/04000000 - Economy, Business and Finance/04016000 - Company Information	14.285714%				
Top/13000000 - Science and Technology/13016000 - Electronics	14.285714%				

-
2. Use the data that appears in the SAS Content Categorization Server Categorization Statistics (Matches) page to gain information about matching categories for the input texts:

Table 4-2: Categorization Statistics (Matches) Information

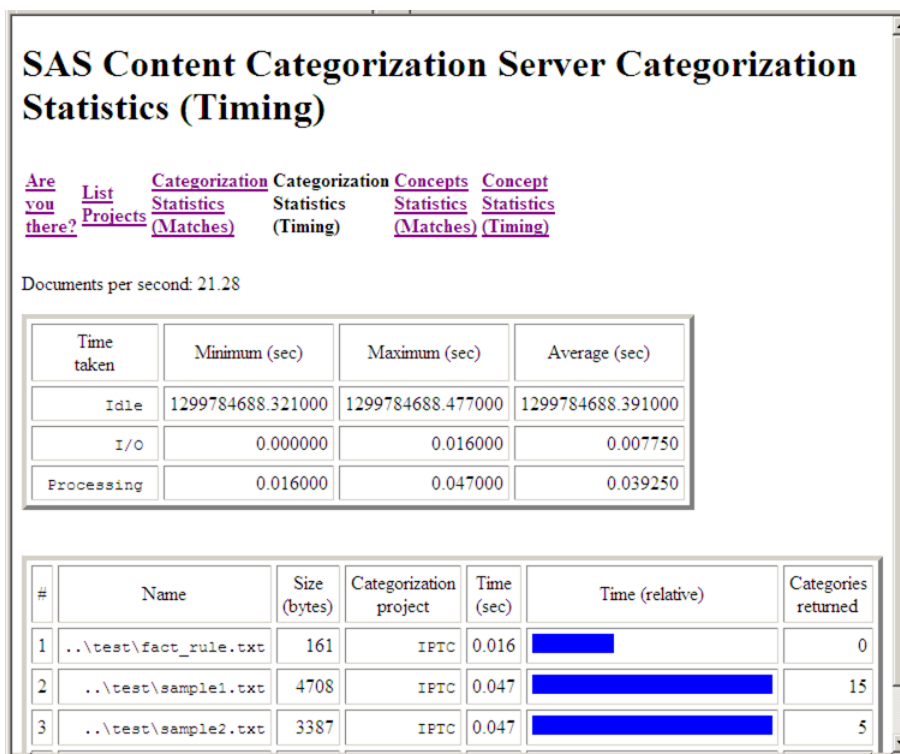
Information Type	Description
Total number of documents	This number represents the total number of documents that match one or more category rules. The example above specifies 7.
Total number of categories with at least one match	This number represents the total number of categories that have one or more matching documents. This example specifies 28 categories have matches.
Category Name	This number includes the full pathname of the category with one or more matching documents.
Percentage of Documents	This number represents the proportion of texts that matched the specified category.
Percentage (relative)	See the bar chart to visually compare the results shown in the Percentage of Documents column.

4.3.5 Use the SAS Content Categorization Server Categorization Statistics (Timing) Page

The SAS Content Categorization Server Categorization Statistics (Timing) page displays data about document processing and timing.

To use the SAS Content Categorization Server Categorization Statistics (Timing) page, complete these steps:

1. Select the **Categorization Statistics (Timing)** link and the SAS Content Categorization Server Categorization Statistics (Timing) page appears. The tables in this screen contain timing information relative to processing the input documents.



2. See the Documents per second statistics to see the total number of documents processed by SAS Content Categorization Server. In the example shown above this number is 21.28.

-
3. Use the information in the first table for the time required to process the input documents:

Table 4-3: First Categorization Timing Table Information

Heading	Description
Time taken	The following types of timing occur with document processing: Idle: The amount of time that SAS Content Categorization Server was not processing documents. I/O: The amount of time that it took to input and output a single text. Processing: The time required to process a document.
Minimum (sec)	The fewest number of seconds used to process any one document.
Maximum (sec)	The highest number of seconds used to process any single text.
Average (sec)	The number of seconds required to process all of the input documents divided by the total number of processed texts.

4. Use the information in the second table to see the number of categories that match each input document:

Table 4-4: Second Categorization Timing Table Information

Heading	Description
#	Each input document incrementally increases by this number, beginning with 1.
Name	The name, if there is one, of the processed document.
Size (bytes)	The size of the processed documents in bytes.
Categorization project	The symbolic name of the categories project that is matched by these categories.

Table 4-4: Second Categorization Timing Table Information

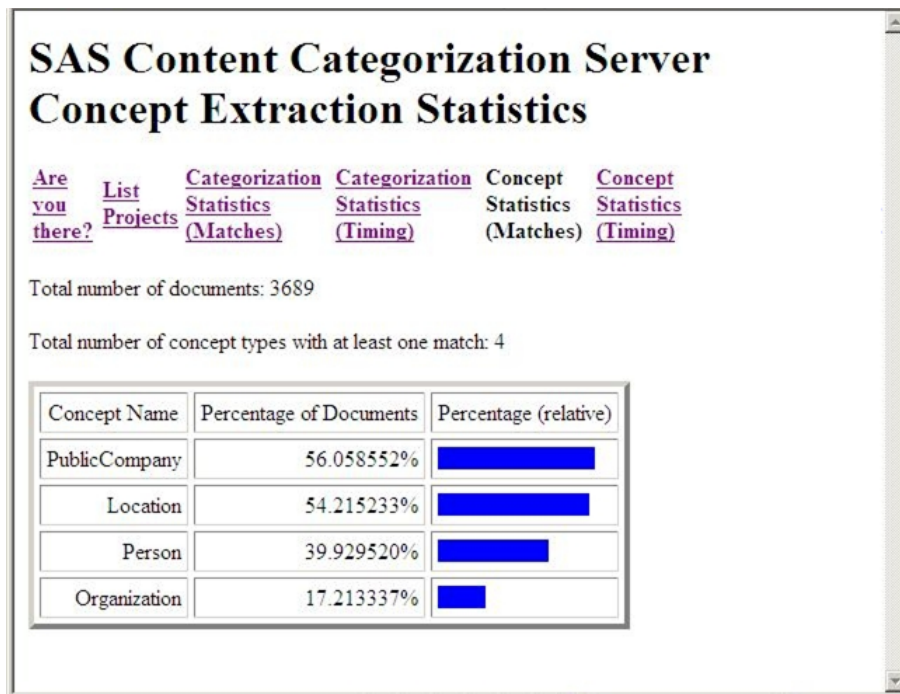
Heading	Description
Time (sec)	The number of seconds that it took to process the specified document.
Time (relative)	An overview of the preceding column Time (sec) . Use this bar chart for comparison purposes.
Categories returned	The number of category rules that this document matches.

4.3.6 Use the SAS Content Categorization Server Concept Extraction Statistics Page

The SAS Content Categorization Server Concept Extraction Statistics page displays data about the concepts that are extracted, or matched, in the input documents. This page displays information about the numbers of documents that are matched and the concepts that they match.

To use the SAS Content Categorization Server Concept Extraction Statistics page, complete these steps:

1. Select **Concept Statistics (Matches)** to see the statistical data compiled for matches in processed documents. The SAS Content Categorization Server Concept Extraction Statistics page appears.



2. Analyze the data that appears in the SAS Content Categorization Server Concept Extraction Statistics page:

Total number of documents

represents the total number of texts that have matched one or more concept definitions. For example, 3689 texts are matched.

Total number of concept types with at least one match

represents to the total number of concepts that have at least one matching document. For example, 4 concepts match at least one input document.

-
3. Evaluate the information that appears in the table:

Concept Name

see the name of the concept that one or more documents match.

Percentage of Documents

use these figures to see the concepts that have the highest percentage of matching documents.

Percentage (relative)

use this bar chart to visually compare the results shown in the **Percentage of Documents** column.

4.3.7 Use the SAS Content Categorization Single User Servers Concept Extraction Statistics (Timing) Page

The SAS Content Categorization Server Concept Extraction Statistics (Timing) page displays data about concepts that are extracted, or matched, in input documents. This page displays information about the numbers of documents that are matched and the concepts that they match.

To use the SAS Content Categorization Server Concept Extraction Statistics (Timing) page, complete these steps:

1. Select the **Concept Statistics (Timing)** link to see these statistics and the SAS Content Categorization Server Concept Extraction Statistics (Timing) page appears.

SAS Content Categorization Server Concept Extraction Statistics (Timing)

[Are you there?](#)
[List Projects](#)
[Categorization Statistics \(Matches\)](#)
[Categorization Statistics \(Timing\)](#)
[Concepts Statistics \(Matches\)](#)
[Concept Statistics \(Timing\)](#)

Documents per second: 86.96

Time taken	Minimum (sec)	Maximum (sec)	Average (sec)
Idle	4.374000	4.484000	4.425000
I/O	0.000000	0.016000	0.011500
Processing	0.000000	0.000000	0.000000

#	Name	Size (bytes)	Concept extraction project	Time (sec)	Time (relative)	Concepts returned
1	..\test\fact_rule.txt	161	Entities	0.000		0
2	..\test\sample1.txt	4708	Entities	0.000		18

2. See processing speed for the documents to the right of **Documents per second**. For example, see 86.96.
3. Analyze the data that appears in the first table. For more information, see Table 4-3 on page 66.

4. Analyze the data that appears in the second table:

Table 4-5: Second Contextual Extraction Timing Table Information

Heading	Description
#	Each input document incrementally increases by this number, beginning with 1.
Name	The name, if there is one, of the processed document.
Size (bytes)	The size of the processed documents in bytes.
Concept extraction project	When more than one concepts project is loaded, this column differentiates between the projects where concepts are extracted.
Time (sec)	The number of seconds that it took to process the specified document.
Time (relative)	An overview of the preceding column Time (sec) . Use this bar chart for comparison purposes.
Concepts returned	The number of concept definitions that this document matches.

Part 3: SAS Document Conversion Server

- Chapter 3: *Processing Various File Types into Plain Text Format on page 37*

Chapter: 5

Processing Various File Types into Plain Text Format

- *Overview of SAS Document Conversion Server*
- *The File Types That SAS Document Conversion Server Can Convert*

5.1 Overview of SAS Document Conversion Server

Use the SAS Document Conversion Server to preprocess different file types into plain text format. For an overview of why this process is often necessary, see Section 1.3 *How Does SAS Content Categorization Single User Servers Work?* on page 7. After you convert your files into plain text, this text can be sent to SAS Content Categorization Server. Use the .mco, .concepts, or .li files uploaded to SAS Content Categorization Server to apply categories and to locate concepts in the input text. For more information about this process, see Figure 1-1 on page 8.

5.2 The File Types That SAS Document Conversion Server Can Convert

SAS Document Conversion Server processes the document formats that are supported by *Apache Tika 1.0*, which is an open-source helper library. See the following table for a list of these file types and their description:

Table 5-1: Supported Document Types

Abbreviation	Format Type	Description
HTML	HyperText Markup Language	Web mark-up language that includes HTML and valid XHTML
XML and XML derived formats	Extensible Markup Language	The custom parsers for some widely used XML vocabularies like XHTML, OOXML, and ODF.
Microsoft Office	OLE 2 Compound Document and Office Open XML (OOXML) formats	The older OLE 2 format was introduced in Microsoft Office version 97 and was the default format until version 2007 and the XML-based OOXML format.
ODF	OpenDocument Format	The default format of the OpenOffice.org office suite and other applications.
PDF	Portable Document Format	A common document format used by Adobe Reader and Acrobat.
EPUB	Electronic Publication Format	The format used for many digital books.
RTF	rich text format	An older document interchange format.
	compression and packaging formats	The various compression and packaging formats such as .zip.
	text formats	The character encoding detection and normalization of plain text docs.
	mbox format	The e-mail messages from the mbox format that is used by many e-mail archives and Unix-style mailboxes.

Part 4: Appendixes

- Appendix A: *Troubleshooting on page 79*
- Appendix B: *Recommended Reading on page 85*
- Appendix C: *Glossary on page 87*

Appendix: A

Troubleshooting

- *Installing SAS Content Categorization Single User Servers*
- *Tips and Guidelines for SAS Content Categorization Server*
- *Using the Configuration File*
- *If SAS Content Categorization Server Does Not Appear to Be Running*

A.1 Installing SAS Content Categorization Single User Servers

A.1.1 The Installer Fails to Copy Files on Windows Vista or Server 2008

If the SAS Content Categorization Single User Servers installers fails to copy files to the program directory on Windows Vista or Server 2008, the user account control might be enabled. Use the following steps even if you are signed in with administrative privileges.

In order to enable the SAS Content Categorization Single User Servers installer to function correctly, complete these steps:

1. Right-click on the SAS_ConCat_Collab_Servers_<arch>_Setup.exe file.
2. Select **Run as...**
3. Enter the administrator information in the Run As window that appears.
4. Click **OK** to close the Run As window.

Note: If you install SAS Content Categorization Single User Servers as an administrator, you are the only person who can uninstall this software.

A.2 Tips and Guidelines for SAS Content Categorization Server

A.2.1 If You Are Unable to Upload a File to SAS Content Categorization Server

Check to see whether there is a `lock` file in the project directory. If there is a `lock` file, delete that file.

A.2.2 Trying to Upload a Project with a Duplicate Name

If you try to upload a project to SAS Content Categorization Server with the same name as a project that has already been uploaded, the existing project is overwritten. This statement is true even when the type of project is different. For example, a `MyProj.desc` file for a `.mco` file can be overwritten by a `MyProj.desc` file for a `.concepts` file.

A.2.3 Noun Phrases Incorrectly Split by an Apostrophe (')

Some noun phrases are incorrectly split by an apostrophe. This is a known issue.

A.2.4 Using Synonym Lists

If you upload a project containing synonyms, the server cannot process documents that are larger than 1 MB in size.

A.3 Using the Configuration File

A.3.1 Trailing Spaces in the Configuration File

If you edit the server configuration file and there is a space at the end of a line that specifies the uploaded filename, the server might not deploy this model. Check all lines and remove all spaces in the configuration file if you find a trailing space.

A.3.2 Modifying the Configuration Backup File

If you installed SAS Content Categorization Server prior to the installation that you now have loaded on your machine, the `server.config.bak` file exists. This backup file overwrites the latest configuration file. For this reason, you might want to manually edit your configuration file. This is especially true if you have a new `setinit` file. For more information, see Chapter 4: *Configuring and Running SAS Content Categorization Server*.

A.4 If SAS Content Categorization Server Does Not Appear to Be Running

A.4.1 Overview of When the Server Does Not Appear to Be Running

If SAS Content Categorization Server does not appear to be running on your machine, access the SAS Content Categorization Server Administration Web Page. You can access this page using the Start menu, see Section 2.5 *Access the Servers* on page 33. Alternatively, open a Web browser and type:

http://<machine_name>:<admin_port>/admin

For example, if the server is running on your local machine, and you selected the default admin port of 6501 during the installation, type:

<http://localhost:6501/admin>

Click the [Are you there?](#) link in the SAS Content Categorization Server Administration Web Page to see whether the server is running.

A.4.2 Checking and Debugging on a Windows Machine

On Windows, the server output is available when you go to **Start --> Control Panel --> Administrative Tools --> Windows Event Viewer**. SAS Content Categorization Server writes data to the Application log. If you can identify the issue in the server output, start a debugging instance of the server to obtain further information.

To start a debugging instance of the server, complete these steps:

1. Shut down the server using the Services window. For more information, see Section 2.3 *Starting and Stopping the Servers on a Windows Machine* on page 27.
2. Open a Windows command prompt window.
3. Enter the following command using the installation path:

```
cd <INSTALL_DIR>\Teragram Catcon Server\  
_catcon_server.exe -server conf\server.config -verbose
```

A.4.3 Checking and Debugging on a UNIX Machine

You can shut down the server and start a debugging instance of the server.

To start a debugging instance, complete these steps:

1. Access a shell window.
2. Enter the following commands:

```
cd /path/to/sas_cc_servers/cc_server
./bin/<arch>/_catcon_server -server conf/server.config
                           -verbose
```

Hint: Replace <arch> with your system's architecture such as linux64.

Appendix: B

Recommended Reading

The following books are recommended:

- *SAS Enterprise Content Categorization Servers: Administrator's Guide*: Install, configure, and use SAS Content Categorization Single User Servers, SAS Content Categorization Collaborative Server, and SAS Document Conversion Server. Upload .li files using this product.
- *SAS Enterprise Content Categorization Studio: User's Guide*: Deploy the collaborative operations and develop complex LITI concepts and their rules for SAS Content Categorization Studio. Use this book after you read *SAS Content Categorization Studio: User's Guide*.
- *SAS Content Categorization Studio: Installation Guide*: Install SAS Content Categorization Studio.
- *SAS Content Categorization Studio: User's Guide*: Create a SAS Content Categorization Studio project, test, and upload the output to SAS Content Categorization Single User Servers.
- *SAS Enterprise Content Categorization Studio: Administrator's Guide*: Install and configure the server used for the collaborative operations available in SAS Enterprise Content Categorization Studio.

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.

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

Web address: 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: C

Glossary

Administration Web Page

moderate the progress and operations of SAS Content Categorization Server.

categorization

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

definition

defines a concept. Sometimes, this manual uses the word *rule* as a synonym for the word *definition*.

document

refers to an input text. Also see *Text*.

plain text

readable textual material that does not require much processing.

rule

defines a category. This term is also used, within this manual, to refer to a concept definition.

string

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

text

form that a written document, or a Web page takes, can be called a text. Also see *Document*.

Index

#	
defined	66, 71

A

admin_port	
defined	41
Administration Web page	
view	58
Are you there?	
defined	59
Average (sec)	
defined	66

B

backupdir	
defined	40
basedir	
defined	40

C

cat_log	
defined	42
statistics	57
cat_log_max_entries	
defined	42
example	51
Categories returned	
defined	67
Categorization project	
defined	66
category	
rules	5
category matches	
log file	42

Category Name	
defined	64
comment character	
configuration file	38
concept	
definitions	5
concept extraction project	
defined	71
concept Name	
defined	69
concept_log	
defined	43
statistics	57
concept_log_max_entries	
defined	43
concepts	
defined	44
Concepts returned	
defined	71
configuration file	
comment character	38
restart program	43
UNIX	39
Windows	39
CPU	12
create_dir	
defined	41
creator	
defined	42

D

Default Category Bias	
defined	62
Default Relevancy Cutoff	
defined	62
definitions	
concept	5
descriptor_dir	
defined	40

do_cat_log_timing	
defined	43
do_concept_log_timing	
defined	43

E

entries	
maximum exceeded	42, 43
entry	
defined	50
ephemeral ports	
defined	54
exceed	
log file limits	51

F

File Name	
defined	62

I

I/O	
defined	66
Idle	
defined	66
installation	
wizard	15
installation guide	
UNIX	27
installation kit contents	11
installation root directory	53
io_log	
defined	42

L

log file	
exceed limits	51
regenerate	51

M

max_doc_size	
defined	41
max_iterations_to_reinitialize	
defined	41
Maximum (sec)	
defined	66
mcats	
defined	44
memory	
project files	51
Minimum (sec)	
defined	66
multiple project files	
specify	51

N

Name	
defined	66, 71
nb_threads	
defined	41
num_cat_logs	
defined	43
num_concept_logs	
defined	43
Number of Rules	
defined	62

O

operating systems	
supported	13

P

Percentage (relative)	
defined	64, 69
Percentage of Documents	
defined	64, 69
persistent_connection	
defined	41
Processing	
defined	66
protocol_version	
defined	43

Q

query_port	
defined	41

R

RAM	12
restart program	
configuration file	43
rules	
category	5

S

SAS Content Categorization Server	
running on Windows	52
UNIX shell command	53
-server configfile	
defined	53
setinit	
defined	40
Size (bytes)	
defined	66, 71
skt_queue_size	
defined	41

stat_cat	
defined	44
statistics	
cat_log	57
concept_log	57
generate	57
Status	
defined	59
Symbolic Name	
defined	62
system configuration	
specified	12

T

Time (relative)	
defined	67, 71
Time (sec)	
defined	71
Time taken	
defined	66
timeout	
defined	41
Total number of categories with at least one match	
defined	64
Total number of concept types with at least one match	
defined	68
Total number of documents	
defined	64, 68
Type	
defined	62

U

UNIX	
configuration file	39
installation guide	27
UNIX shell command	
running the server	53
Uptime	
defined	60

user	
defined	42

W

Windows	
configuration file	39
installation guide	14

