THE POWER TO KNOW.

SAS® Content Categorization Server 5.2

Administrator's Guide

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2011. *SAS® Content Categorization Server 5.2: Administrator's Guide*. Cary, NC: SAS Institute Inc.

SAS® Content Categorization Server 5.2: Administrator's Guide

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** or call 1-800-727-3228.

Contents

About This Book

Audience

SAS Content Categorization Server is designed for the following users:

- Creators who upload binary files for new project to SAS Content Categorization Server.
- Users who deploy the Web administrator interface to see the matching statistics for categories and concepts that are applied to input documents.

Prerequisites

Here are the prerequisites for using SAS Content Categorization Server:

- Load SAS Content Categorization Server onto your machine.
- Create one of more projects in SAS Content Categorization Studio.

Conventions

This manual uses the following typographical conventions:

Convention	Description
`TGM_ROOT`	The root directory where SAS Content Categorization Server is installed, typically the following: **Windows:** `C:\Program Files\Teragram\Teragram Catcon Server\` **UNIX:** `/opt/sas_catcon_server_linux64`
Next button	The labels for user interface controls are shown in a bold, sans-serif font.
`www.sas.com`	The hypertext links are shown in a light blue, fixed-width font, and are underlined.

What's New in SAS Content Categorization Server 5.2

The new features in SAS Content Categorization Server include the following:

- Most of the configuration work that was recently required to upload a binary project to SAS Content Categorization Server is now unnecessary. The creator has administrative permissions to perform this task.
- SAS Contextual Extraction Studio is now supported in SAS Content Categorization Server.
- SAS licensing replaces Teragram license

Chapter: 1

About SAS Content Categorization Server

1.1 What is SAS Content Categorization Server?

In most organizations it is necessary to obtain information about, and from, data that is created internally and externally. SAS Content Categorization Server applies categorization and concept extraction to input documents. Categorization and concept extraction are applied using the category rules and concept definitions that your organization developed in SAS Content Categorization Studio. You can also upload the binary files created by SAS Contextual Extraction Studio to SAS Content Categorization Server. These files enable SAS Content Categorization Server to apply contextual extraction concept definitions to input documents.

Use the output from SAS Content Categorization Server with the custom application of your choice. For example, use a custom application to insert appropriate indexing tags into your documents.

SAS Content Categorization Server is available as a UNIX server or as a Windows service. You can manage SAS Content Categorization Server by using standard UNIX or Windows service management tools. For example, use the Web administrator interface that comes with SAS Content Categorization Server to see the matching data in real time.

Easy project upload

The administrator with the creator role can upload binary files from SAS Content Categorization Studio and SAS Contextual Extraction Studio using a window in SAS Content Categorization Studio.

Easy configuration

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

Web administrator interface

Administrators can use the Web administrator interface to stop, start, and to see data on matched rules, definitions, documents, and timing.

1.2 Benefits of Using SAS Content Categorization Server

SAS Content Categorization Server provides users with the following benefits:

Automatically locate matching documents

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

Gain real-time knowledge of matches

The administrative Web interface enables you to see the statistics generated by matching 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.

1.3 How Does SAS Content Categorization Server Work?

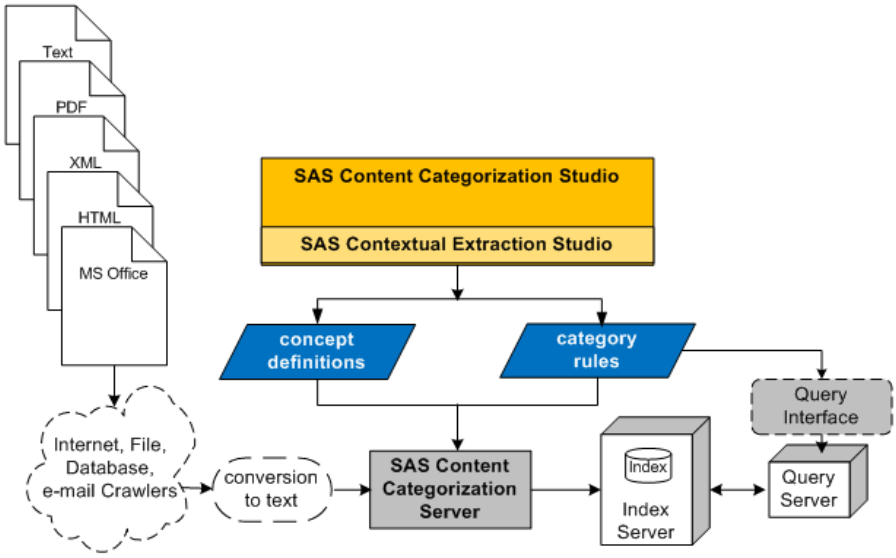
SAS Content Categorization Server is an application that administrators and other users use to automatically classify documents and to extract concepts from the input texts. SAS Content Categorization Server applies the rules and definitions in the form of binary files. These rules and definitions are written in SAS Content Categorization Studio and SAS Contextual Extraction Studio.

1.4 About the Architecture

SAS Content Categorization Server provides rapid, run-time categorization and concept extraction for documents collected from your corporate intranet or the Internet. This application runs on a server, deploying the binary files created by SAS Content Categorization Studio and SAS Contextual Extraction Studio. SAS Content Categorization Server automates the application of the compiled category rules and concept definitions in the binary files.

Administrators can also use the administrative Web interface to view various types of statistical reports on the matched categories and concepts.

Figure 1-1 SAS Content Categorization Server Architecture



Chapter: 2

Installing SAS Content Categorization Server

2.1 Before You Install SAS Content Categorization Server

2.1.1 Overview of Installation

This chapter explains the hardware requirements and the installation process for SAS Content Categorization Server. This chapter also explains how to specify the path to the SAS license.

The SAS Content Categorization Server installation kit for Windows (`SAS_ConCat_Server_Setup.exe`) contains all of the components required to install (and uninstall) SAS Content Categorization Server.

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.

Configure the system where you install SAS Content Categorization Server in accordance with the recommended system requirements.

2.1.2 Prerequisite System Requirements

Configure the machine where you install SAS Content Categorization Server 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 Server:

Table 2-1: Supported Operating Systems

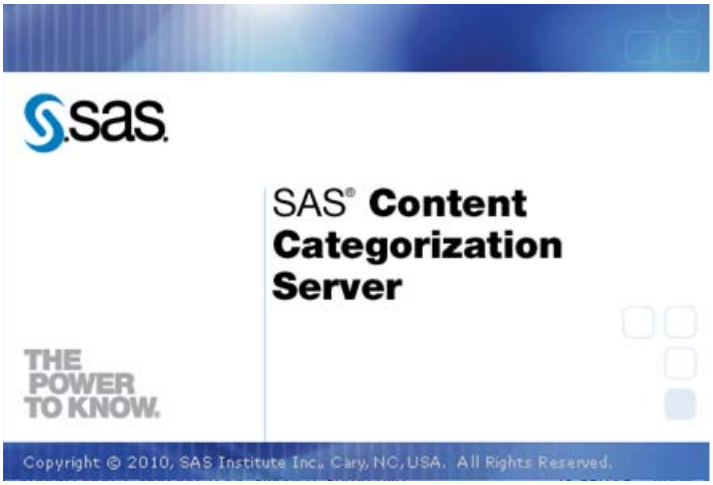
Operating System	Platform
Linux, (Red Hat 7.x, 8, 9, Fedora 1-3, RHEL 2.1 and higher), SUSE	x86, x86-64
IBM AIX	PPC
HP-UX	PA-RISC
Sun Solaris (32-bit)	SPARC
Sun Solaris (64-bit)	UltraSPARC, x86-64
Windows	x86, x86-64

2.2 Installing SAS Content Categorization Server

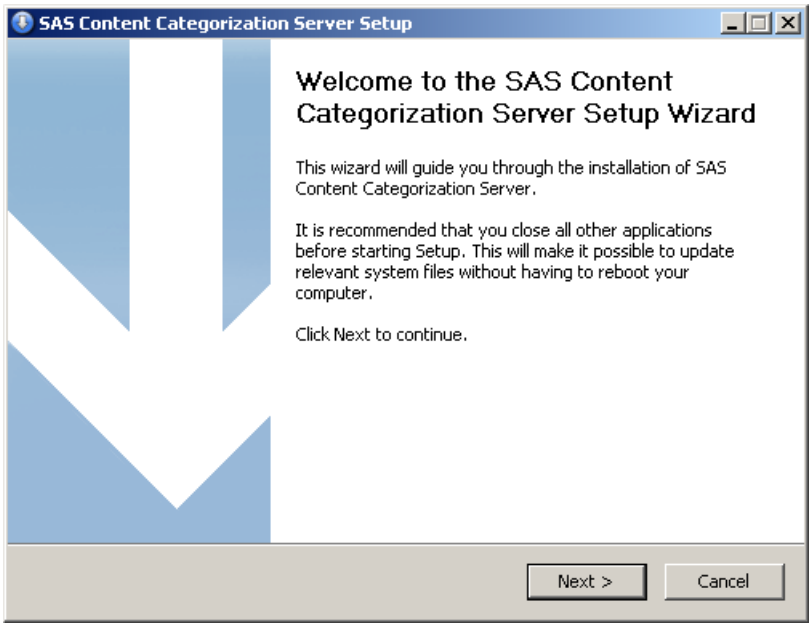
2.2.1 Install on Windows

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

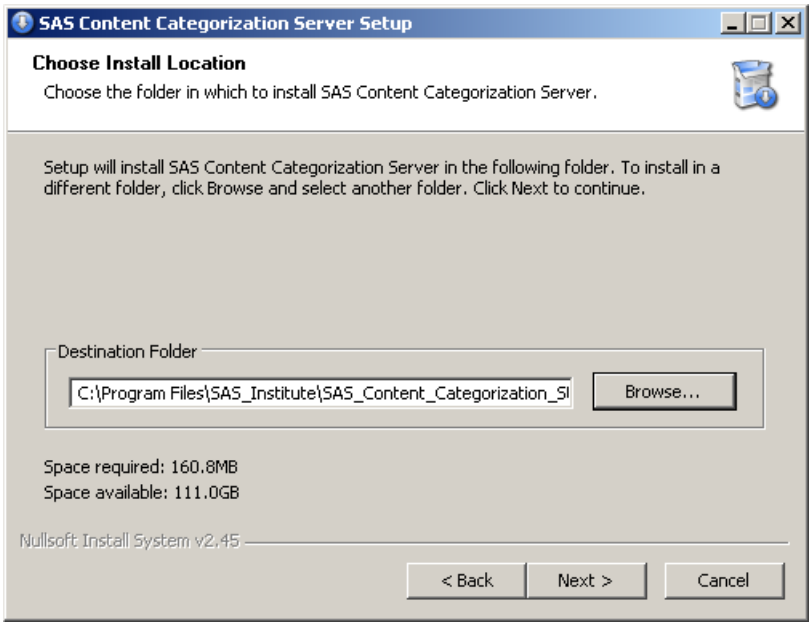
1. Double-click `SAS_ConCat_Server_Setup.exe` and the installation wizard appears.



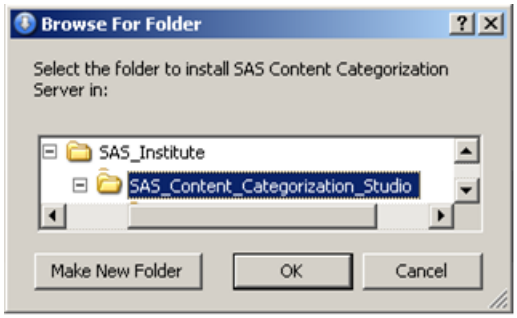
The Welcome page appears.



2. Click **Next** and the Choose Install Location page appears.



3. (Optional) Click **Browse** and the Browse For Folder dialog box appears.



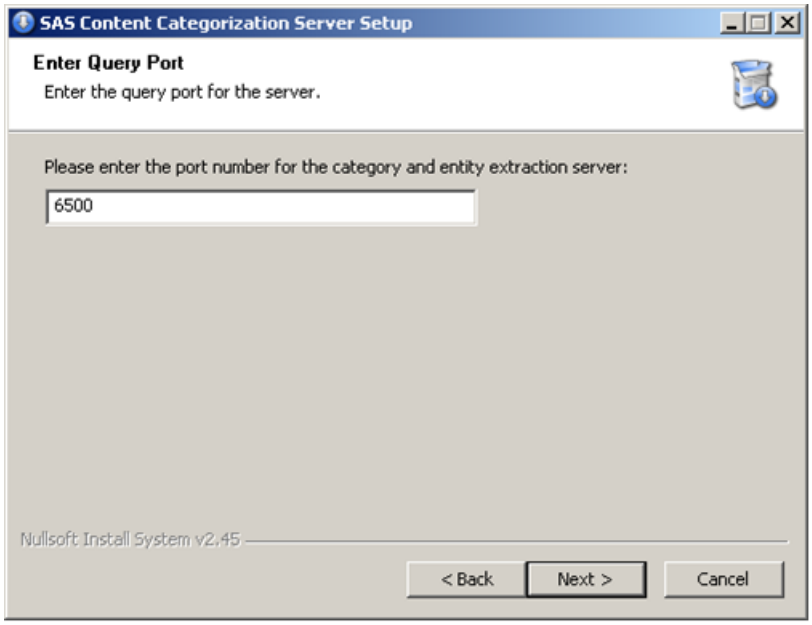
   a. (Optional) Select a different installation folder.

   b. (Optional) Click **Make New Folder**.

   c. Click **OK**.

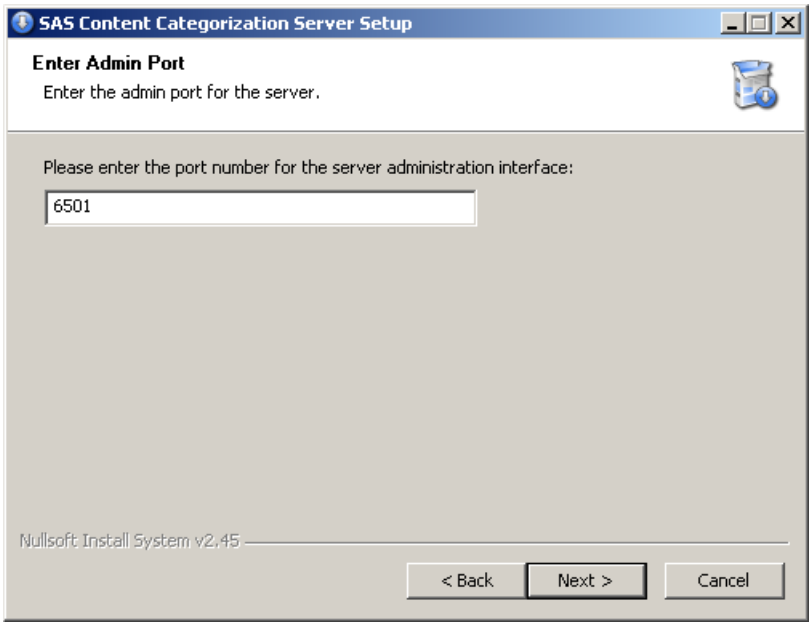
4. Compare **Space required** with **Space available** in the Choose Install Location page to ensure that there is enough room on your hard drive for the application.

5. Click **Next** and the Enter Query Port page appears.



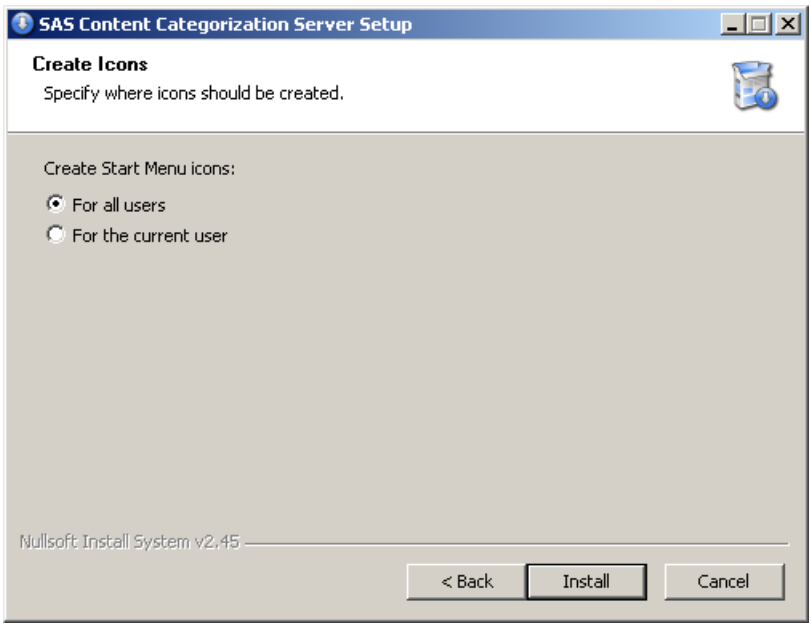
6. (Optional) Enter a new Query Port Number in the **Please enter the port number for the category and entity extraction server**.

7. Click **Next** and the Enter Admin Port page appears.

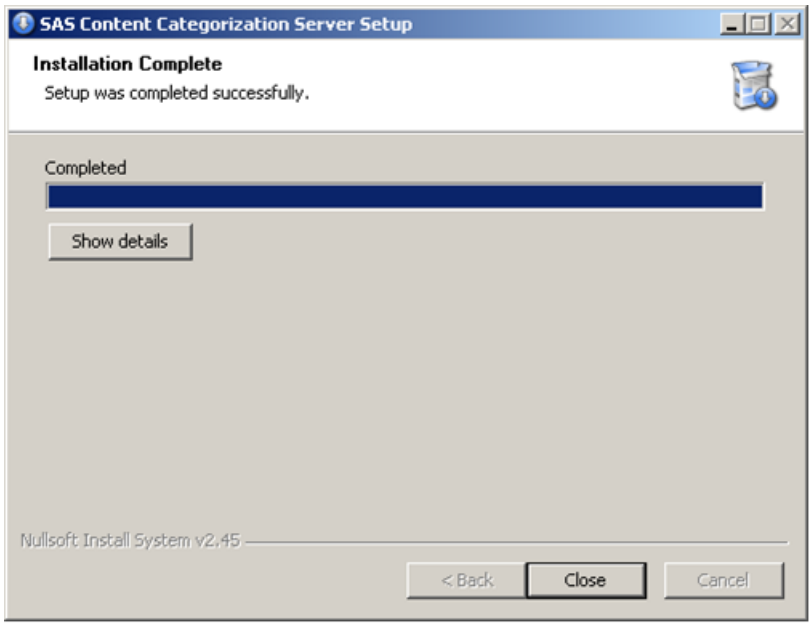


8. (Optional) Enter a new admin port number in the **Please enter the port number for the server administration interface field**.

9. Click **Next**. The Create Icons page appears:



10. Click **Install** and the Installation Complete page appears.



11. Click **Close**.

2.2.2 Install on UNIX

SAS Content Categorization Server is distributed on UNIX systems as a `tar` archive. 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 `-d` switch on the `gzip` command decompresses the distribution file (compressed to save space) in preparation for the expansion of the archive `tar` file. One example of the `.gz` filename is `sas_cc_server_linux64.tar.gz`. 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.

2.2.3 Specify the Path to the SAS License

This 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 `cntcatsvr.sas` to the directory on your hard drive that is specified in the `server.config` file. The setinit file can also be saved in any folder, as long as you specify the path relative to the base directory in the `server.config` file.

The following example of a `server.config` file provides an example of the path:

Example 2-1: Server Configuration File

```
basedir=C:\Program Files\Teragram\Teragram Catcon
Server\
backupdir=backup
setinit=cntcatsvr.sas
descriptor_dir=descriptors
create_dir=data
user=pat:pat1
creator=pd:p1d
mcat=data/English.mco:IPTC
concepts=data/English.concepts:Entities
query_port=6500
admin_port=6501
skt_queue_size=10
nb_threads=4
persistent_connection=0
timeout=60000000
max_iterations_to_reinitialize=5000
```

For more information about the configuration file for SAS Content Categorization Server, see Chapter 3: *Configuring and Running the Server*.

2.3 Uninstall on Windows

To uninstall the SAS Content Categorization Server software on a Microsoft Windows system go to **Start --> Programs --> SAS Content Categorization Server --> Uninstall SAS Content Categorization Server**.

Chapter: 3

Configuring and Running the Server

- *Configuration Overview*
- *Server Configuration File*
- *Setting the Path to the SAS License*
- *Specifying Project Users and Creators*
- *Add a Project*
- *Using the Data Directory*
- *Using the Descriptor Directory*
- *Using cat_log and concept_log Files*
- *Specifying Multiple Project Files*
- *Specifying Input Documents*
- *Running the Server*
- *Optimizing Performance on a Client Windows Machine*

3.1 Configuration Overview

You configure and run SAS Content Categorization Server when you specify directives in the configuration file (`server.config`). These directives enable you to specify creators who are administrators that 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.

Before you process documents using SAS Content Categorization Server, configure the server using a text file that contains key - value pair

assignments. Use 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 binary files are stored in the `descriptors` directory after you install SAS Content Categorization Studio. See the following examples:

Example 3-1: ITPC.desc

```
type=mcat
path=data\English.mco
name=IPTC
```

Example 3-2: Entities.desc

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

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

3.2 Server Configuration File

3.2.1 Windows Configuration File

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

```
C:\Program Files\Teragram\Teragram Catcon
                  Server\conf\server.config
```

To see this file, select **Programs --> SAS Content Categorization Server --> 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\Teragram Catcon
            Server\
backupdir=backup
setinit=cntcatsvr.sas
descriptor_dir=descriptors
create_dir=data
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
```

Example 3-4: Sample Configuration File for UNIX

```
basedir=/opt/sas_catcon_server_linux64
backupdir=backup
setinit=cntcatsvr.sas
descriptor_dir=descriptors
create_dir=data
query_port=6500
admin_port=6501
skt_queue_size=10
nb_threads=4
persistent_connection=0
timeout=60000000
max_iterations_to_reinitialize=5000
```

On a UNIX system, you name and choose where to store the SAS Content Categorization Server configuration file. There is no name or location for this file. To load the configuration file, run SAS Content Categorization Server and provide the full pathname with the `-server` switch.

3.2.3 The Directives

Use the directives to modify the server configuration file. Directives such as `descriptor_dir` that were relative to `basedir=` in earlier versions, are now 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 examples:

```
backupdir=c:\backups
setinit=/home/sasuser/cntcatsvr.sas
```

The directives for the configuration file are described in the table below:

Table 3-1: Configurable Directives

Directive	Description
`basedir`	Specifies the path to the project binaries, backup directory, and so on.
`backupdir`	Specifies the directory where the backup binaries are stored. When a categorization (`.mco`), concepts (`.concepts`), or contextual extraction (`.li`) 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.
`setinit`	Specifies the SAS license file for SAS Content Categorization Server in a path that is relative to `basedir`. For example, this license file could be `cntcatsvr.sas`, or another `name` that you choose. This SAS installation data file (SID) is in the Software Order E-mail (SOE) that you received.
`descriptor_dir`	Contains text files with information about the projects that are loaded on SAS Content Categorization Server. The path represented in `basedir` is prepended to the `descriptor_dir`. If this path is not available, `descriptor_dir` is treated as an absolute path.

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. The path represented in basedir is prepended to the create_dir. If this path is not available, descriptor_dir is treated as an absolute path.
query_port	Categorization and concept extraction services are available on this TCP port number. The clients connect to this port on the server host.
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. The 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. The maximum number of clients that can be served is the number of threads on the server.
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 by the categorizer. Larger texts are truncated.

Table 3-1: Configurable Directives (Continued)

Directive	Description

`xml_weight_file`	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 `xml_weight_file` is `field: weight` for each line in the file. You could specify the following:
	`title:3`
	`body:1.5`
	In this example, if a match is located in the `body` field of an XML document, the match counts 1.5 times toward the relevancy score. However, a match for the `title` field is multiplied by three.

`user`	Specifies a user name and password. Users specify these strings to upload the binary files directly from SAS Content Categorization Studio to SAS Content Categorization Server. This statement is true if a version of the file that is uploaded already exists. The format is `User Name:password`. For example, see `Alice:wonderland`, for the user *Alice* with the password *wonderland*. **Tip**: This directive can be specified multiple times.
`creator`	Specifies a user name and password for the users who have the privileges necessary to upload new projects and reload existing projects. **Tip**: This directive can be specified multiple times.
`io_log`	Generates a detailed input and output log file while performing the categorization operation. This log includes timestamps. The value is the filename. **Tip**: This directive only applies to categories.
`cat_log`	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 filename. For more information, see Section 3.8 *Using cat_log and concept_log Files* on page 32.
`cat_log_max_entries`	Specifies the maximum number of entries allowed in each `cat_log` file. For more information, see Section 3.8 *Using cat_log and concept_log Files* on page 32.

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 3.8 *Using cat_log and concept_log Files* on page 32.
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 the 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 3.8 *Using cat_log and concept_log Files* on page 32.
concept_log_max _entries	Specifies the maximum number of entries allowed in each concept_log file. For more information, see Section 3.8 *Using cat_log and concept_log Files* on page 32.
num_concept_logs	Specifies the maximum number of concept log files to create. For more information, see Section 3.8 *Using cat_log and concept_log Files* on page 32.
do_concept_log _timing	Provides 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 to take effect.

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
`mcat`	Specifies either a relative, or an absolute, path to a categorization binary file (`.mco`). The symbolic name for the categories project is specified after a colon (`:`).
`concepts`	Specifies either a relative or an absolute path to a concepts binary file (`.concepts`). The symbolic name for the concepts project is specified after a colon (`:`).
`liti`	Specifies either a relative or an absolute path to a concept extraction binary file (`.li`). This `.li` file also includes the `.concepts` file for any concepts referenced by contextual extraction concepts. The symbolic name for the liti project is specified after a colon (`:`).
`stat_cat`	Specifies either a relative or an absolute path to a statistical categorizer binary file (`.st.cat`). The symbolic name is specified after the colon (`:`).
Tips: These directives can be specified multiple times. If `basedir` is specified, these paths are relative to this directory. If this directory is not specified, these paths are absolute.	

3.3 Setting the Path to the SAS License

Before you run the server, specify the SAS license file or the server does not start. This SAS installation data file (SID) for SAS Content Categorization Server is located in the Software Order E-mail (SOE) that you received. Save the license file, `cntcatsvr.sas`, to the installation directory of SAS Content Categorization Server.

If you convert from a Teragram to a SAS license, insert the following line. (The configuration files for updated applications include this line.)

```
setinit=<path>
```

Specify a relative path to the `cntcatsvr.sas` file, unless this file is located in the installation directory.

3.4 Specifying Project Users and Creators

You can specify one class of users and one class of administrators in the SAS Content Categorization Server configuration file:

User

can reload a binary file for a project that is already uploaded to SAS Content Categorization Server by a creator. 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. This class of administrators creates new projects. This upload process automatically adds the appropriate line to SAS Content Categorization Server.

Using either directive, specify the name and password in the **Username** and **Password** fields. These specifications apply to any of the upload windows in SAS Content Categorization Studio.

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

Example 3-5: User and Creator Directives

```
user=user1:pw1
user=user2:pw2
creator=creator1:pw3
creator=creator2:pw4
```

All duplicate usernames are ignored after the first instance. For example, if you specify:

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

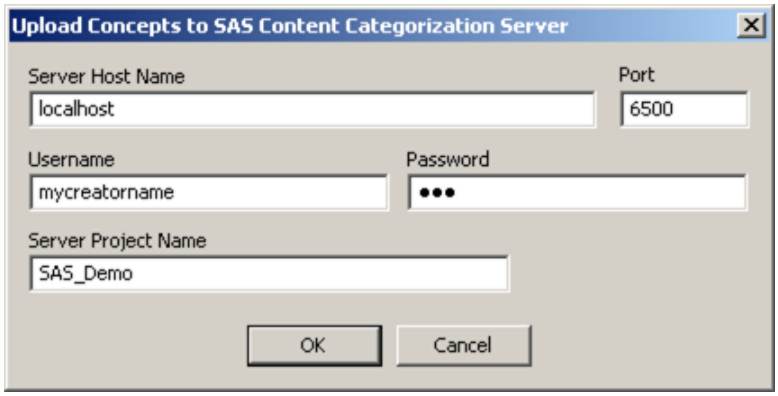
In this example, Joe has creator permissions. If you specify `-verbose`, the server emits a warning for duplicate or conflicting entries.

3.5 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, 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. It is also unnecessary to copy the binary files to the data directory.

To add a `.concepts` project developed in SAS Content Categorization Studio to SAS Content Categorization Server, complete these steps. Make appropriate changes for other binary file types.

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



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

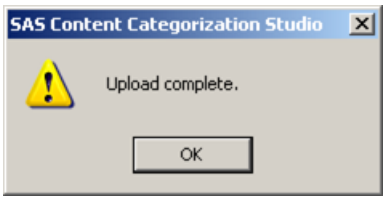
Note: The value for the **Port** field should be the value specified in the `query_port` directive of the `server.config` file, not the `admin_port` field.

---

2. Type your creator name into the **Username** field.

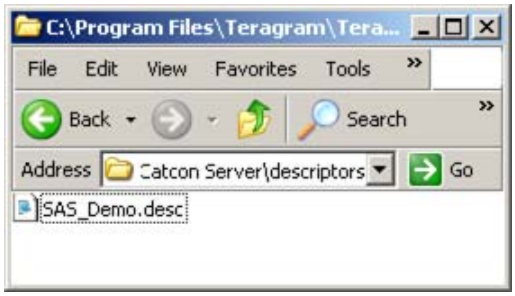
3. Type your creator password into the **Password** field.
4. Type the symbolic name for the file that you choose to upload into the **Server Project Name** field.
5. Click **OK**. A SAS Content Categorization Studio status window appears.

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

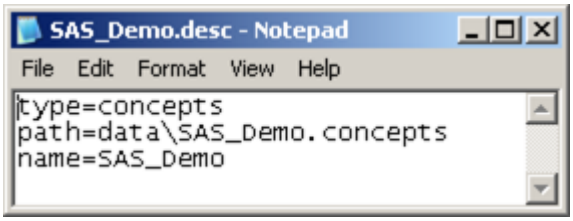


7. Click **OK** and the project is listed in the `data` folder.

8. The project is also listed in the descriptors folder as a `<project name>.desc` file. For example, see `SAS_Demo.desc`:



9. Double-click the `.desc` file to open and read the file contents. For more information, see Section 3.7 *Using the Descriptor Directory* on page 31.



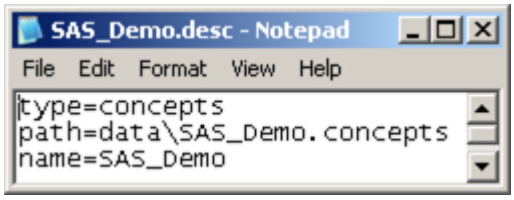
3.6 Using the Data Directory

The `data` 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, `English.concepts` and `English.mco` might be the names of these two files. See the following example:

3.7 Using the Descriptor Directory

The descriptor directory (`descriptor_dir`) contains information about each project that is uploaded to SAS Content Categorization Server. A descriptor file named `<project name>.desc` is automatically created in `C:\Program Files\Teragram\Teragram CatCon Server\descriptors`. 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. See the following example of a descriptor file:



The format for this file is displayed below:

`type=[mcat|concepts|liti]`

specifies the type of project that is uploaded. You can upload a categories (`mcat`), concepts (`concepts`), fact extraction (`lei`), or contextual extraction (`liti`) 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`, `concepts` (`lei` for Teragram legacy customers only), and `liti` directives can co-exist with the `descriptor_dir` directive.

3.8 Using cat_log and concept_log Files

This section provides examples of the 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 Web administration tool, specify the appropriate log files in your `server.config` file.

For example, if an input text matches one category in each of two projects, two entries are created. If however, if 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 3.9 *Specifying Multiple Project Files* on page 34.)

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.

3.9 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. For example, the first project specified by `mcat=` and the first project specified by `concepts=` are the default categorization and concepts files, respectively.

These project names are also used by the APIs for SAS Content Categorization Server.

3.10 Specifying Input Documents

Use the client APIs to send documents to the server. These client APIs are available in the C, Perl, Python, C#, and Java.

3.11 Running the Server

3.11.1 Run the Server on Windows

To start the SAS Content Categorization Server service, go to **Start --> Programs --> SAS Content Categorization Server --> Start SAS Content Categorization Server**.

The configuration file that was created during installation is automatically loaded.

3.11.2 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.

3.12 Optimizing Performance on a Client Windows Machine

3.12.1 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.

3.12.2 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 3.12.3 *Adjust the TCP Time Wait State* below and Section 3.12.4 *Reset Ephemeral Ports* on page 36, reboot your machine.

3.12.3 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 typing `TcpTimedWaitDelay`.
4. Double-click on the new `TcpTimedWaitDelay` value.
5. Select `Decimal` as the base, and enter `15` for the value data.

3.12.4 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 typing `MaxUserPort`.
4. Double-click on the new `MaxUserPort` value.
5. Select `Decimal` as the base, and enter `65534` for the value data.

Chapter: 4
Web Administration

4.1 Overview of Web Administration

Use the SAS Content Categorization Server administrative Web interface, 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.

The administrative Web interface 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.
- Reload your projects.

Note: The statistics generated for category matches and concept extraction are returned only if the `cat_log` and `concept_log` options are specified in the server configuration file.

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

4.2 Open the Administrative Web Interface

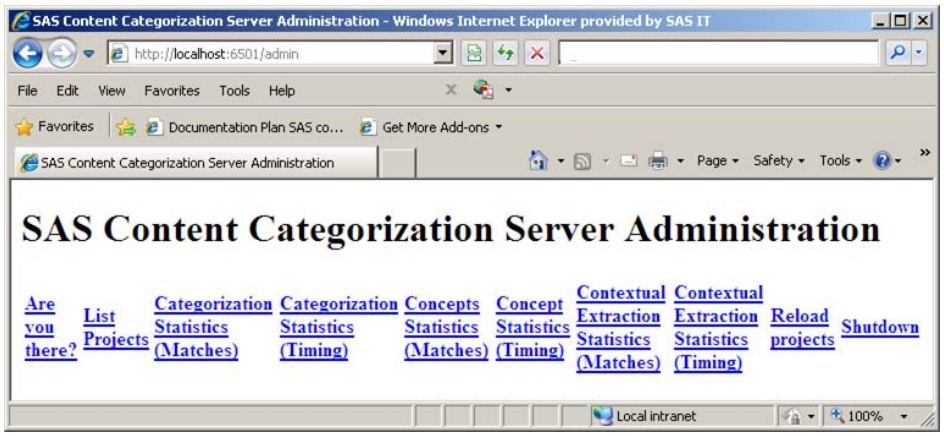
After you launch the SAS Content Categorization Server, you can deploy the administrative Web interface:

To launch SAS Content Categorization Server, go to **Start —> Programs —> SAS Content Categorization Server —> Start SAS Content Categorization Server**.

Open your Web browser and type the following address:

```
http://localhost:6501/admin
```

The SAS Content Categorization Server - Administration interface appears:



4.3 Using the Administrative Web Interface

4.3.1 Overview of the Administrative Web Interface

Use the Administration Web 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.

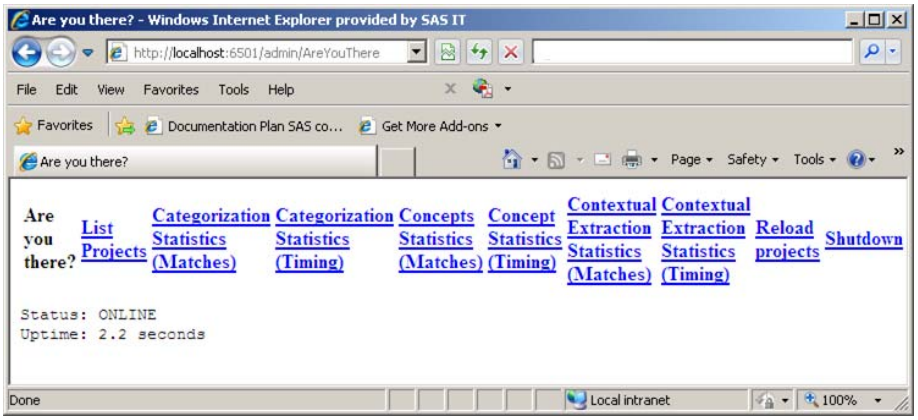
To run, access, and use the various management operations that are accessed through links on this page, use the following sections. Each operation has an assigned section in this chapter. The sections are ordered to match the links starting on the left side of the page and progressing to the right side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

Uptime

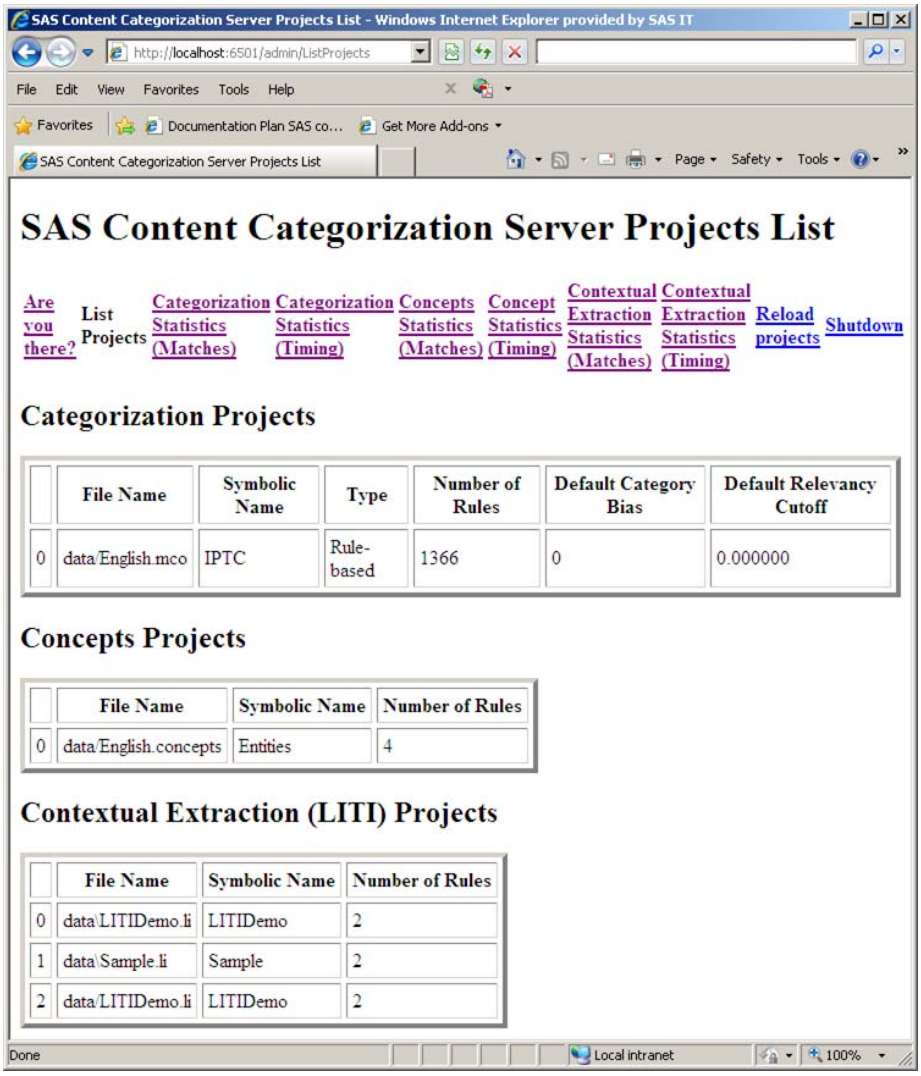
displays the length of time that SAS Content Categorization Server has been running, or when the process was restarted. In the example above, the time is `2.2` seconds.

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.



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

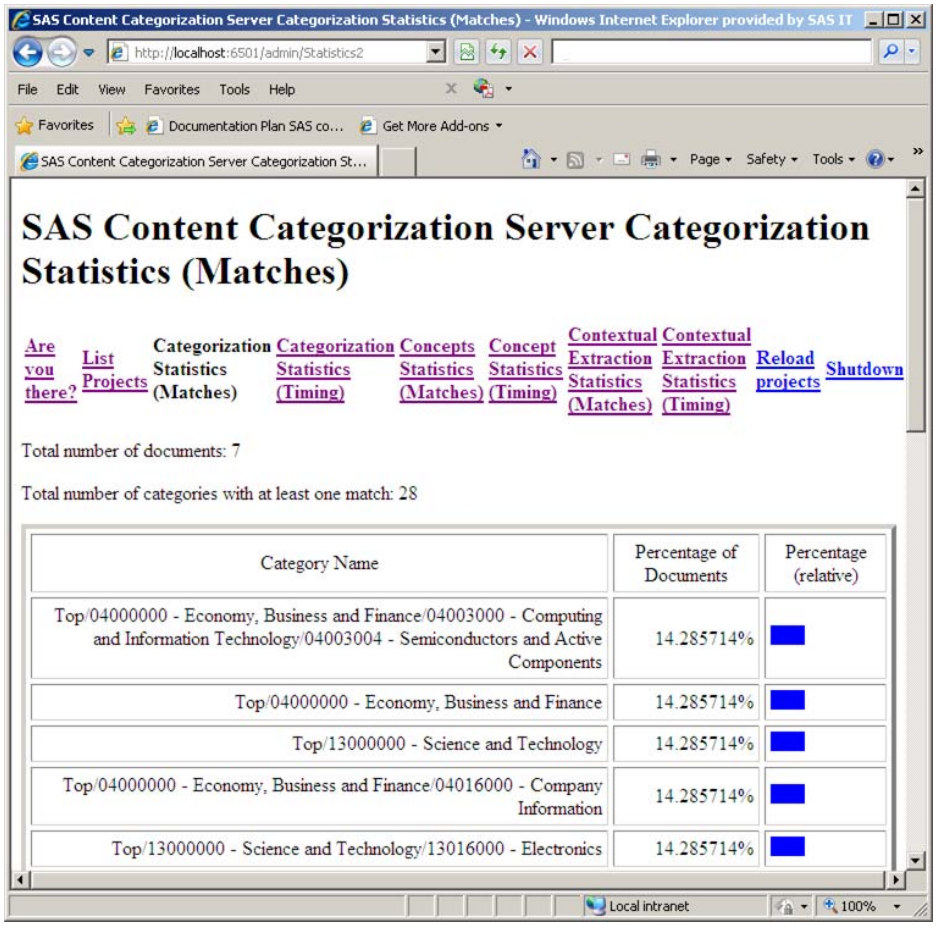
Heading	Description
Categorization Projects	
File Name	Specifies the name and location of the project file that was exported from SAS Content Categorization Studio. For example, the `English.mco` file located in the data folder.
Symbolic Name	Specifies the name of the project that you enter into the 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.
For contextual extraction Projects, see the information that is specified in the table cells beneath Concepts Projects above.	

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.



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.

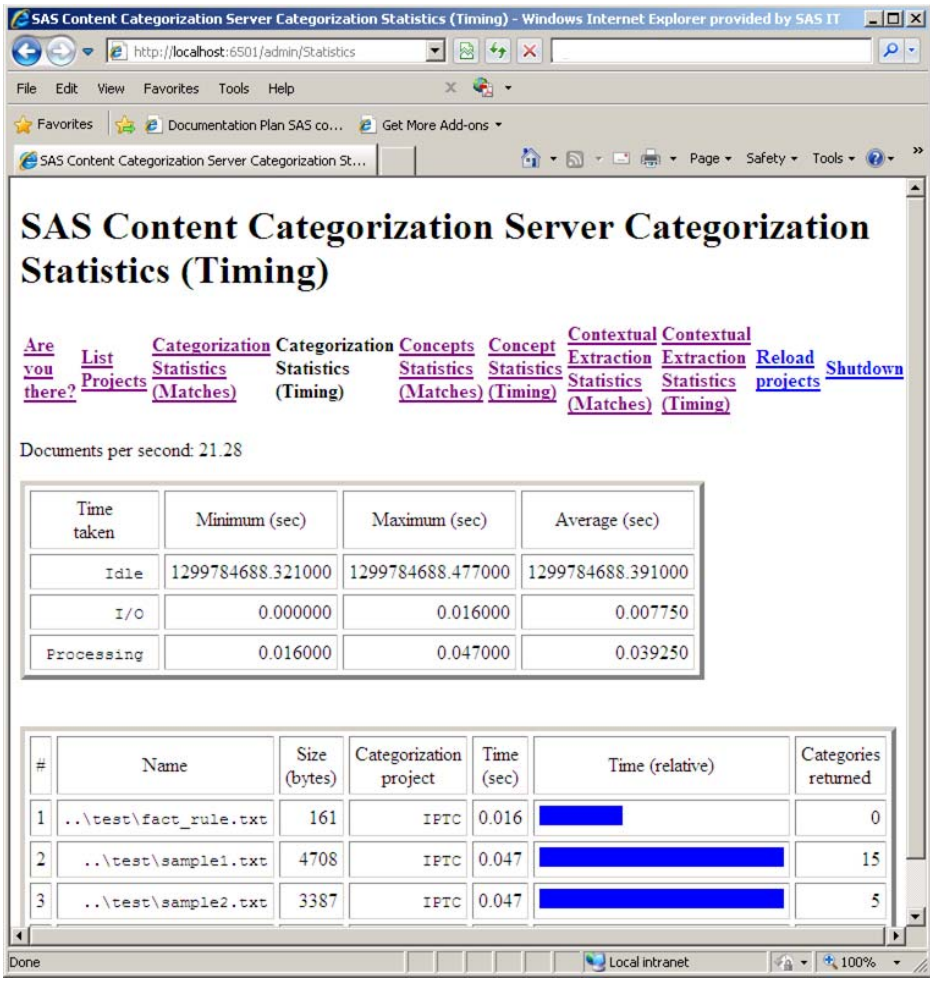
4.3.5 Use the SAS Content Categorization Server Categorization Statistics (Timing) Page

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 about 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: This is the amount of time that SAS Content Categorization Server was not processing documents. I/O: This is the amount of time that it took to input and output a single text. Processing: This is the time required to process a document.
Minimum (sec)	This is fewest number of seconds used to process any one document.
Maximum (sec)	This is 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. **Note**: The projects loaded are determined by the server configuration file. For more information, see Chapter 3: *Configuring and Running the Server*.

Table 4-4: Second Categorization Timing Table Information

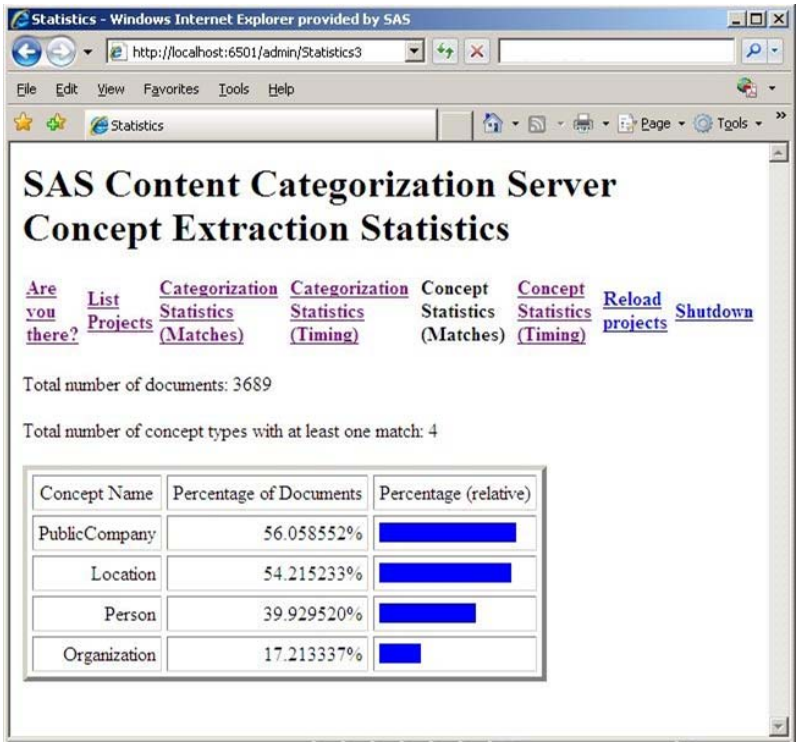
Heading	Description
Time (sec)	The number of seconds that it took to process the specified document.
Time (relative)	An overview of the preceding column **Time (sec)**. Use this bar chart for comparison purposes.
Categories returned	The number of category rules that this document matches.

4.3.6 Use the SAS Content Categorization Server Concept Extraction Statistics (Matches) Page

The SAS Content Categorization Server Concept Extraction Statistics (Matches) 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 (Matches) page, complete these steps:

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



2. Analyze the data that appears in the SAS Content Categorization Concept Extraction Statistics (Matches) 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

   refers 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)

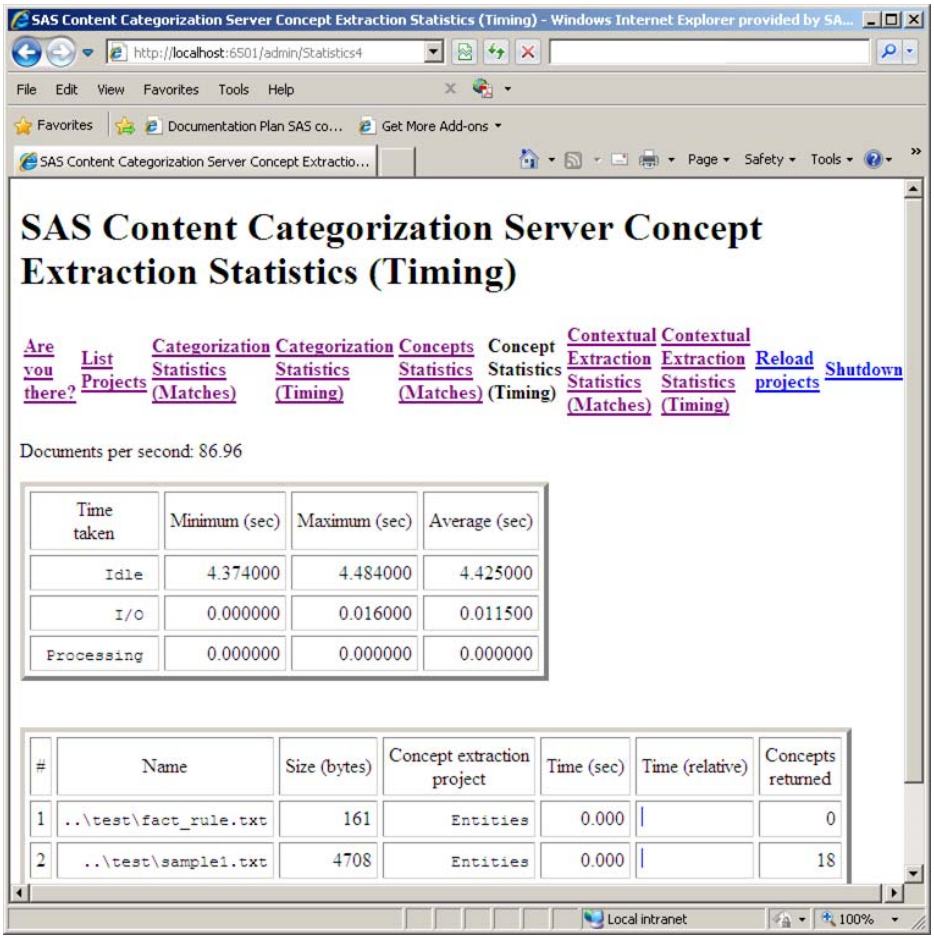
use this bar chart to visually compare the results shown in the **Percentage of Documents** column.

4.3.7 Use the SAS Content Categorization Server 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.



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 48.

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

Table 4-5: Second Contextual Extraction Timing Table Information

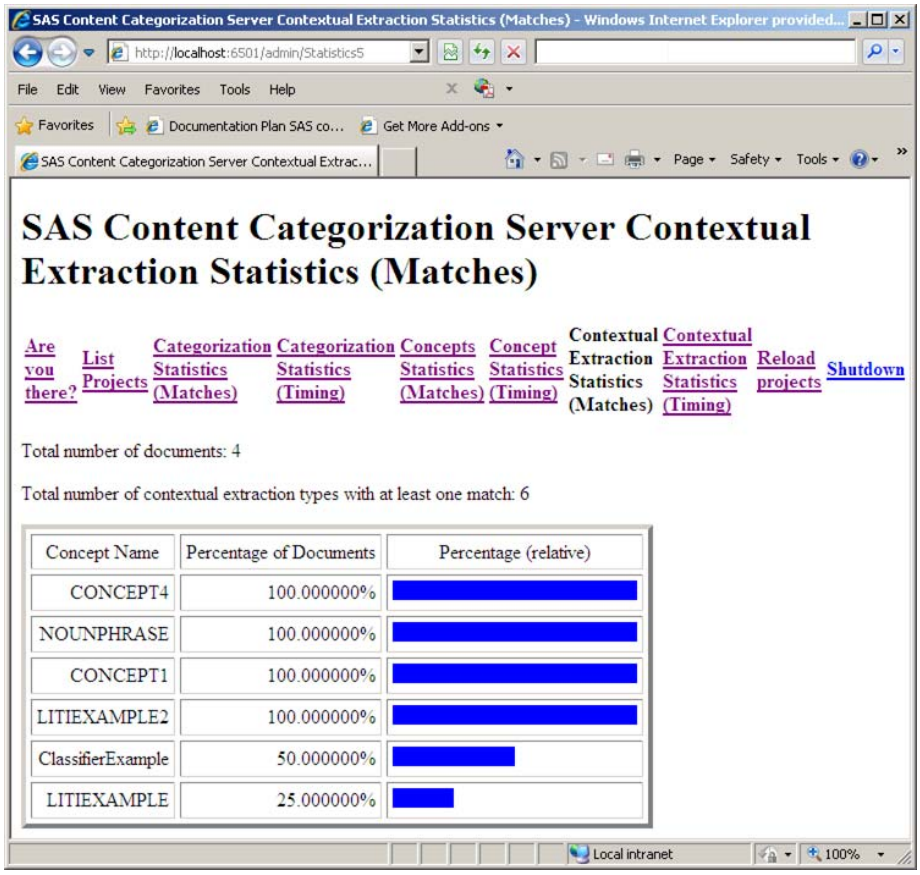
Heading	Description
#	Each input document incrementally increases by this number, beginning with 1.
Name	The name, if there is one, of the processed document.
Size (bytes)	The size of the processed documents in bytes.
Concept extraction project	When more than one concepts project is loaded, this column 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.

4.3.8 Use the SAS Content Categorization Server Contextual Extraction Statistics (Matches) Page

The SAS Content Categorization Server Contextual Extraction Statistics (Matches) page displays data about the contextual extraction concepts 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 (Matches) page, complete these steps:

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



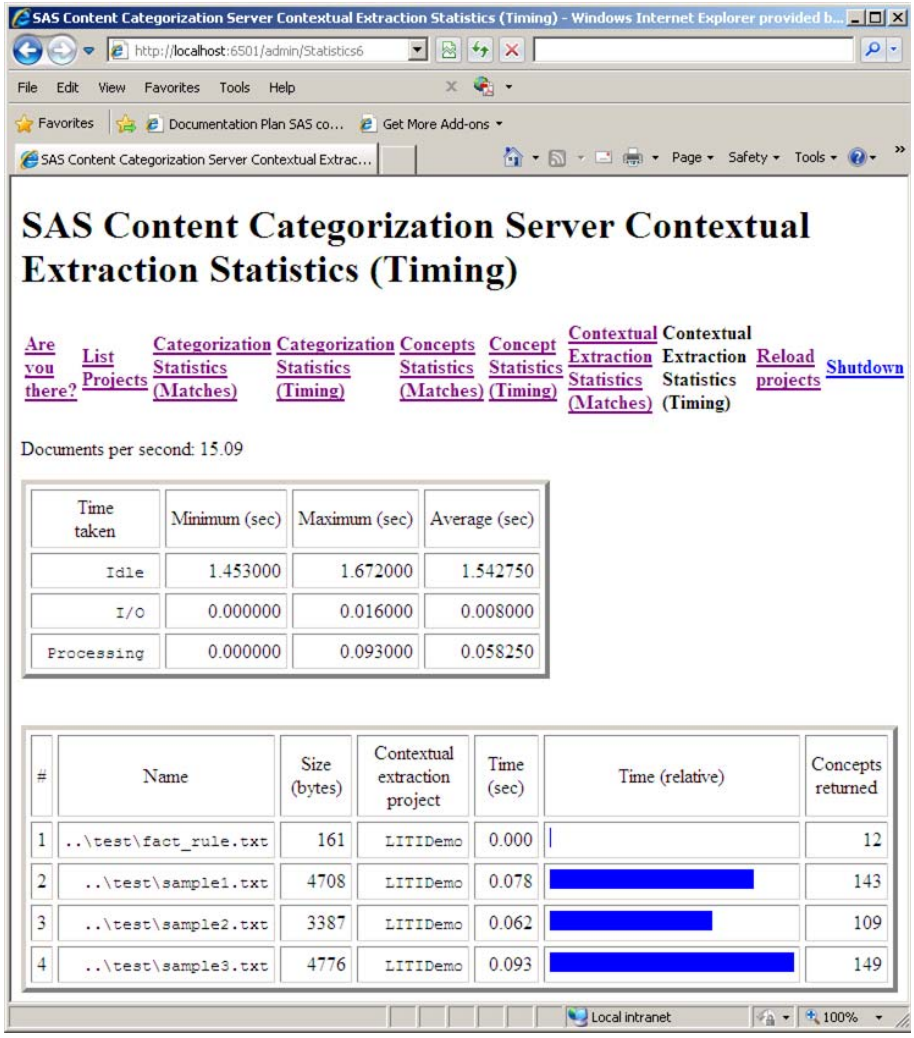
2. See the total number of documents processed to the right of **Total number of documents**. For example, see 4.
3. See the **Total number of contextual extraction types with at least one match**. For example, see 6.
4. Analyze the data that appear in the first table. For more information see Section 4.3.4 *Use the SAS Content Categorization Server Categorization Statistics (Matches) Page* on page 45.

4.3.9 Use the SAS Content Categorization Server Contextual Extraction Statistics (Timing) Page

The SAS Content Categorization Server Contextual Extraction Statistics (Timing) page displays data about the time that it takes to process 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 (Timing) page, complete these steps:

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



2. See the total number of documents processed per second to the right of **Documents per second**. For example, see `15.09`.

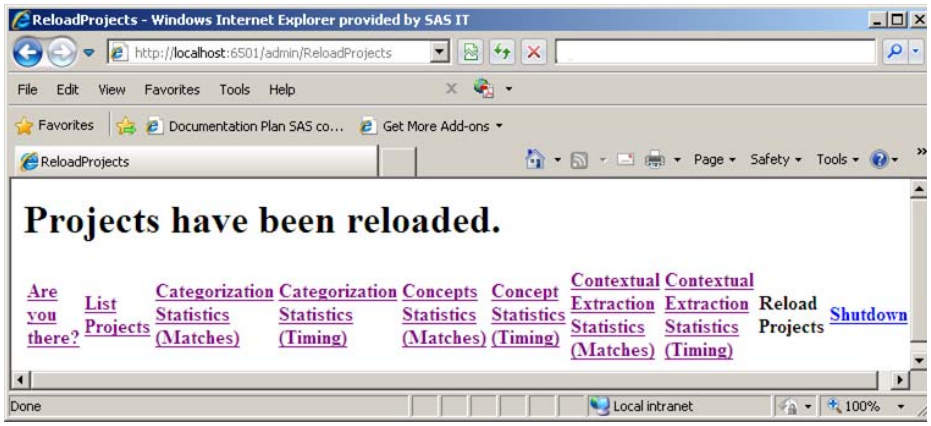
3. Analyze the data that appear in the first table. For more information see Section 4.3.6 *Use the SAS Content Categorization Server Concept Extraction Statistics (Matches) Page* on page 49.

4.3.10 Use the Reload Operation

To upload changed projects, use the Reload Projects operation. This process overwrites the existing project with the new project.

To reload your projects, complete these steps:

1. Click **Reload Projects**.



2. Check the status of this operation in the **Projects have been reloaded** page that appears.

4.3.11 Use the Shutdown Operation

The SAS Content Categorization Server Concept Shutdown page displays information about closing the SAS Content Categorization Server program.

To shutdown SAS Content Categorization Server Concept Extraction Statistics and to see the status of this operation, complete these steps:

1. Click **Shutdown**.

2. See the status of this operation to the right of **Info**. For example, see `Shutting down now`.

Appendixes

Appendix: A
Recommended Reading

The following books are recommended as companion guides:

- *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 Content Categorization Studio: Quick Start Guide*: Advanced users can learn how to expeditiously set up a SAS Content Categorization Studio project.
- *SAS Content Categorization Collaborative Server: User's Guide*: Enable multiple subject matter experts to work together on one SAS Content Categorization Studio project.
- *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.
- 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.

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: B
Glossary

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

document

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

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

D

E

F

I

L

M

N

O

P

Q

R

S

T

U

W