IBM Text Analyzer Business Component

Installation Guide for UNIX

Version 1.1

... A member of the WebSphere Business Components family

Before using this information and the product it supports, be sure to read the general information under “Notices” on page 10.

Second Edition (March 2001)

This edition applies to version 1.1 of IBM Text Analyzer Business Component (part number 20P4393 or 20P4395), and to all subsequent releases and modifications until otherwise indicated in new editions. Make sure you are using the correct edition for the level of the product.

Corrections and suggestions for future revisions of this document are appreciated. Mail your comments to:

IBM Canada Ltd. Laboratory
Information Development
2G/KB7/1150/TOR
1150 Eglinton Avenue East
Toronto, Ontario, M3C 1H7
Canada

When you send information to IBM, you grant to IBM a nonexclusive right to use or distribute the information in any way they believe appropriate without incurring any obligation to you.

Contents

General installation and configuration procedures

This installation guide describes how to install and set up Text Analyzer on a UNIX system. Although the development environment must be Microsoft® Windows NT, this guide describes how to set up a UNIX production environment.

The IBM Text Analyzer (TA) Business Component is installed as two distinct and separately deployable subcomponents:

- Categorization Engine (CE)
    - installed and deployed as a WSBC Version 1.1 Advanced Component (TA AC)
    - temporarily deployable as a Java API
- Training Tool (TT)
    - installed and deployed as a Java API

In this installation guide, installation means installing the required prerequisite and product files; deployment means configuring the product and environment to make the product usable.

Installation prerequisites

The following are the prerequisites for installing TA files.

Software

The Text Analyzer installation procedure has an dependency on WebSphere Business Components Studio. You must install WSBC Studio before you will be able to install any Text Analyzer component files.

The following are the supported UNIX operating systems:

- AIX (4.3.3 and ptf 9) or Sun Solaris 7

Hardware

The Text Analyzer installation procedure does not have any hardware prerequisites for installation beyond those required by WSBC Studio. A full installation of Text Analyzer, including documentation, requires 73539Kb of available hard drive space.

Production deployment prerequisites

The WSBC Studio AC Deployment Tool must be installed and deployable in the production environment for managing the Text Analyzer AC. Refer to the *WSBC Studio Installation Guide for UNIX* for information on installing and using this tool.

To deploy the Categorization Engine as a TA AC in a production environment, WebSphere Application Server must be installed and configured as described in the *WSBC Studio Installation Guide for UNIX*. Advanced Component services must then be installed and deployed in the production environment as described in that installation guide.

WebSphere Application Server provides JDK 1.2.2. If this is installed in your environment then you may use this JDK for the following purposes.

- To temporarily deploy the Categorization Engine as a Java API rather than as an AC
- To deploy the Training Tool Java API

If you are deploying either of the above APIs and do not have JDK 1.2.2 (or later), then you should obtain JDK from the Sun web site and follow the installation instructions provided with the product.

The Text Analyzer environment must provide the Sun Java API for XML Parsing Version 1.0.1 Final Release. This is available by download from java.sun.com/xml and should be installed by following its installation instructions. (At the time of this writing, this version is being replaced by Java API for XML Processing Version 1.1. The Early Access releases are NOT recommended.) JAXP must be made accessible to the application server instance where you run the Text Analyzer AC.

Text Analyzer Categorization Engine and Training Tool API have a runtime dependency on IBM XML Parser for Java (XML4J) Version 2.0.15 or later. If you have WebSphere Application Server 3.5 installed in your production environment, then you already have the required version of XML4J (the default location is <WebSphereRoot>/AppServer/lib/xml4j.jar. You may obtain the parser separately from the IBM alphaWorks Web site (The current version is 3.1.1). XML4J must be made accessible from the application server instance where you run the Text Analyzer AC.

Installing Text Analyzer files

To install WSBC Text Analyzer Business Component 1.1 on UNIX, first install it on Microsoft® Windows NT using the instructions in the *IBM Text Analyzer Business Component Installation Guide* on the CD. Then follow the instructions in the *IBM WSBC Studio Installation Guide for UNIX* to install both the Studio and the Text Analyzer components on a UNIX machine. For convenience, these instructions are repeated here.

1. Zip up the <*WSBC root*> directory, using a tool that provides directory structure support.
2. Create a `wsbc` directory on UNIX (for example: `mkdir /usr/wsbc`). This directory will be the $WSBC_HOME directory.
3. FTP the NT installation zip file to that directory.
4. Extract the NT installation zip file using directory structure support.
5. Copy the UNIXScripts.zip file to the same $WSBC_HOME directory, and extract the zip file using directory structure support. This zip file contains Korn shell scripts for setting up the database and environment, as well as scripts for running the WSBC AC tools.

The following tables describe the Text Analyzer files and their locations after installation. The default installation folder `$WSBC_HOME/ACFeatures/Components/TextAnalyzer` contains all Text Analyzer files. The xml4j.jar (XML4J) and parser.jar (JAXP) files are required to complete an installation as described in "Production deployment prerequisites."

Text Analyzer folder

File name	Description
Events.xsd	Text Analyzer AC event data definition schema
Functions.xsd	Text Analyzer AC functions definition schema
TextAnalyzerAC.jar	Text Analyzer AC implementation
TextAnalyzerACEJB.jar	Text Analyzer deployable EJB implementation
TextAnalyzerACEJBClient.jar	Interface stubs for remote Text Analyzer clients
TextAnalyzerACEJBDeployed.jar	WebSphere deployed Text Analyzer AC EJB
TextAnalyzerBase.jar	Training Tool and Categorization Engine

TextAnalyzer/is	
File name	**Description**
InterfaceModels.xsd	WSBC TextAnalyzerAC type definition schema

TextAnalyzer/LPResources/AIX/bin	
File name	**Description**
libeflnl27x.so	AIX language processing executable
libPoeJNI.so	AIX JNI for language processing

TextAnalyzer/LPResources/Solaris/bin	
File name	**Description**
libeflnl27s.so	Solaris language processing executable
libPoeJNI.so	Solaris JNI for language processing

TextAnalyzer/LPResources/WIN/bin	
File name	**Description**
eflnl27w.dll	Windows language processing executable
poejni.dll	Window language processing JNI

TextAnalyzer/LPResources/dictionaries	
File name	**Description**
files named *.dic and *.abr	Language-specific dictionaries and abbreviations. See online Text Analyzer references for the complete list of supported languages.

TextAnalyzer/rose/TextAnalyzerAC/Interface	
File name	**Description**
TextAnalyzerAC.cat	Model of the Text Analyzer AC and its interface

TextAnalyzer/TTGUIV1R1	
File name	**Description**
data.xml	Training sample template
projectini.xml	Project configuration template
projectList.xml	List of existing projects
TTGUI.bat	V1R1 GUI launcher

TextAnalyzer/TTGUIV1R1/Samples/FullReuters-21578	
File name	**Description**
data.xml	Training sample
projectini.xml	Training configuration
rules.xml	Categorization rule set

TextAnalyzer/TTGUIV1R1/Samples/MiniReuters-21578	
File name	**Description**
data.xml	Training sample
projectini.xml	Training configuration
rules.xml	Categorization rule set

Configuring the NT installed files

The `$WSBC_HOME/WSBCconfig.ini` file must be changed for UNIX-specific information. Change the `WSInstallBase` line to point to where the WebSphere Application Server is installed, and change the `InstallBase` line to point to the $WSBC_HOME directory. The result might look like the following:

```
. . .
WSInstallBase=/usr/WebSphere/AppServer
...
InstallBase=/usr/wsbc
. . .
```

Configuring the UNIX script files

Edit the `$WSBC_HOME/bin/setwsbcenv.sh` file to specify the installation locations of WebSphere, DB2®, and WebSphere Business Components.

```
. . .
WAS_HOME=/usr/WebSphere/AppServer
. . .
WSBC_HOME=/usr/WebSphere/AppServer/wsbc
. . .
DB2_HOME=/home/db2inst1/sqllib
. . .
```

This script is used by other scripts to set the WebSphere and the WSBC home directory. The `$WSBC_HOME/bin/setwsbcenv.sh` script can also be used to set up the client classpath.

Setting up the production environment

Deploying the Training Tool in a production environment.

Install the following files to your Training Tool production environment where they will be accessible to your customized GUI (or some other front-end) implementation. (See the online help documentation for building a customized GUI and for instructions on deploying and configuring the V1R1 Training Tool reference GUI.)

- TextAnalyzerBase.jar
- eflnl27w.dll
- POEJNI.dll
- Language dictionaries as required (*.dic and *.abr files)
  Note: The language dictionary requirements are solely determined by the languages actually used for text analysis. See the online help documentation for more information on the use of language dictionaries. There is no harm in installing all dictionaries if space is available.

Configure your Training Tool front-end implementation to use the TT API.

Configure the TT API settings according to the location of the above listed files (if required).

Installing a deployable Text Analyzer AC in a production environment

The Text Analyzer AC is installed as an application-server-deployable EJB and its supporting resources. Install the following files to your production environment where they will be accessible to your application server instance:

- TextAnalyzerBase.jar
- TextAnalyzerAC.jar
- TextAnalyzerACEJBDeployed.jar
- XML4J.jar
- efln27w.dll
- POEJNI.dll
- Language dictionaries as required (*.dic and *.abr files) by the Categorization Engine configurations

Notes:

1. XML4J is supplied with WebSphere Application Server as WebSphereRoot>/AppServer/lib/xml4j.jar and should be configured as described in the "Production environment deployment prerequisites" section.
2. It is recommended that the DLL files be placed in <WebSphereRoot>/AppServer/bin/.
3. The language dictionary requirements are solely determined by the languages actually used for text analysis. See the online help documentation for more information on the use of language dictionaries. There is no harm in installing all dictionaries if space is available.

Deploying a TA AC instance using WebSphere Application Server

Before deploying Text Analyzer you should deploy AC services. (See "Deploying Advanced Component Services on WebSphere Application Server" in the *IBM WebSphere Business Components Studio Installation Guid*e.)

Configuring WebSphere Application Server to deploy Text Analyzer AC instances

1. Start the WebSphere Application Server AdminServer service.
2. Open the WebSphere Administrator's Console.
3. Select the topology view and then navigate to and select the server instance on which you want to deploy the Text Analyzer AC. A pane will appear on the right side.
4. Enter the following command line arguments and values into the **Command line arguments** text field on the **General** tab of the configuration.
    - -mx128m
    - -classpath

        Add the following paths to the classpath value.

        - <your full path>/TextAnalyzerBase.jar
        - <your full path>/TextAnalyzerAC.jar
        - <your full path>/TextAnalyzerACEJBDeployed.jar

    Notes:

    1. The memory allocation option may be adjusted as required to suit your needs.
    2. Be sure to specify the full file system paths for the .jar files in the classpath argument and to separate the entries with a semicolon.
    3. If the classpath argument already exists then you should simply append the additional classpath values to the existing classpath value.
5. On the **Advanced** tab, take note of the value of the **Transaction Timeout** property and the value of the **Transaction Inactivity Timeout** property. If you experience timeout errors (often reported as CORBA errors) you should increase these values. The tested values resolving timeout errors is 600,000, but lower values may be adequate for your environment.
6. Navigate up to the node (machine) where your server is located. In the **Dependent classpath** field enter the same classpath value as found in Step 4.

Deploying an instance of the Text Analyzer AC using WebSphere Application Server

1. Start WebSphere Application Server service.
2. Start the WebSphere Application Server Administrator's Console.
3. Select the topology view.

4. Navigate down the server tree to your EJB container.

5. Open the **Create EJB dialog** with a right-click on the container to open the context menu, select **Create** from the menu, and finally select **Enterprise Bean** from the submenu.

6. Click on the **Browse** button and navigate to the location of the TextAnalyzerACEJBDeployed.jar file and select the file to view the EJBs it contains.

7. Select the com.ibm.wsbc.ac.textAnalyzerInterface/TextAnalyzerInterface.ser EJB. If you are prompted to enable work load management, select **No**. (The navigator closes and you are returned to the Create EJB dialog which now has fields filled in.)

8. Ensure that the EJB has a unique JNDI HomeName within the container. Click **Edit** on the Create EJB dialog to open the Deployment Properties dialog. Edit the **JNDI HomeName** field; the default name is the fully qualified package name followed by the Session Bean identifier followed by the instance name (com/ibm/wsbc/ac/textanalyzer/TextAnalyzerInterface). Any name may be used provided that it is unique. Click **OK** to close the Deployment Properties Dialog.

9. Register the Text Analyzer AC with the WebSphere Application Server JNDI.

    a. Run the WSBC AC Deployment Tool and select **WebSphere** as the EJB Server

    b. Load the TextAnalyzerAC.jar file.

    c. Deploy the TextAnalyzer AC instance.
    Note: Refer to the *WSBC Studio Installation Guide* and the WSBC Studio online help documentation for AC Deployment Tool details.

Installing and deploying a Categorization Engine configuration in a production environment

The instructions below describe how to install and deploy CE configurations once they have been produced by the Training Tool (The TA AC Sample provides a ready sample CE configuration on Windows NT). CE configuration file sets may be installed and deployed as they become available. A Text Analyzer AC's Categorization Engine uses a CE configuration produced by the Training Tool. You may install the TA AC Sample CE configuration on Windows NT if you do not have any of your own that are ready.

Installing a CE configuration

Install the CE configuration files in a location local to your application server instance:

- Training configuration file produced by the Training Tool
- Rule set file produced by the Training Tool

One consideration is that your TA AC client code must supply the location of the CE configuration files when it initializes the TA AC.

Deploying a CE configuration

The CE configuration must be modified for the local environment. Specifically, the training configuration file must be modified to indicate the location of the dictionary files.

1. Locate the training configuration file (or files) that are installed.
2. In each training configuration file that you deploy, locate the following XML element:

```
<ini>
        <section name="LanguageConfiguration">
                    <setting active="yes" attribute="PathForDictionary">
                            <--Insert your dictionaries path here-->
                    </setting>
        </section>
</ini>
```

3. Change the element data to be the file system path to the installed language dictionaries. This path must be the full absolute path.

Notices

IBM may not offer the products, services, or features discussed in this document in all countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Canada Ltd.,
Department 071,
1150 Eglinton Avenue East
Toronto, Ontario, M3C 1H7
Canada

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement, or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems.

Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks and service marks

The following terms are trademarks of International Business Machines Corporation in the United States, or other countries, or both:

AIX
CICS
DB2
DB2 Universal Database
e-business
IBM
LANDP
MQSeries
OS/2 Warp
OS/390
RS/6000
SanFrancisco
VisualAge
Visual Banker
WebSphere

Lotus, Domino, Lotus Notes, and Notes Mail are trademarks of the Lotus Development Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

MMX, Pentium, and ProShare are trademarks or registered trademarks of Intel Corporation in the United States, other countries, or both.

UNIX is a registered trademark in the United States, other countries, or both and is licensed exclusively through X/Open Company Limited.

Rational Rose is a registered trademark of Rational Software Corporation.

Other company, product, and service names may be trademarks or service marks of others.

End of document