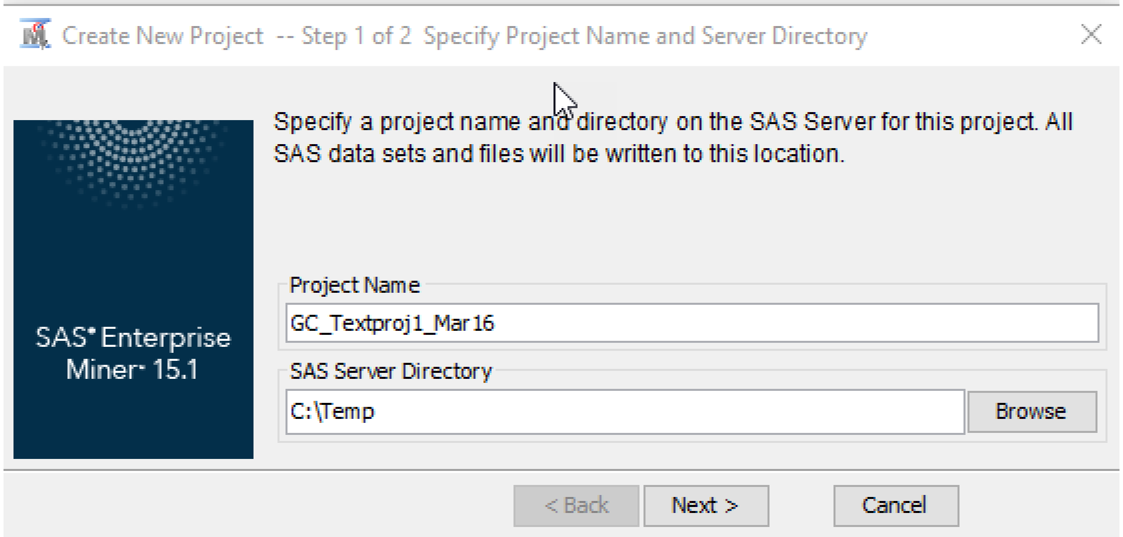
Handout: How to get text data in SAS EM

Bringing Textual Data into Enterprise Miner: How to Import PDF Files

In this example, let us first see how we can use the Text Import node to convert some SAS Global Forum Conference proceedings that are available in PDF format to a SAS data set. Proceedings from 11 SAS Global Forums are available in the folder D:\workshop\winsas\BTASM\Source_PDF.

1. Start SAS Enterprise Miner and select **New Project**. Give the project a suitable name and use an appropriate path for the SAS Server directory.



Create New Project -- Step 1 of 2 Specify Project Name and Server Directory

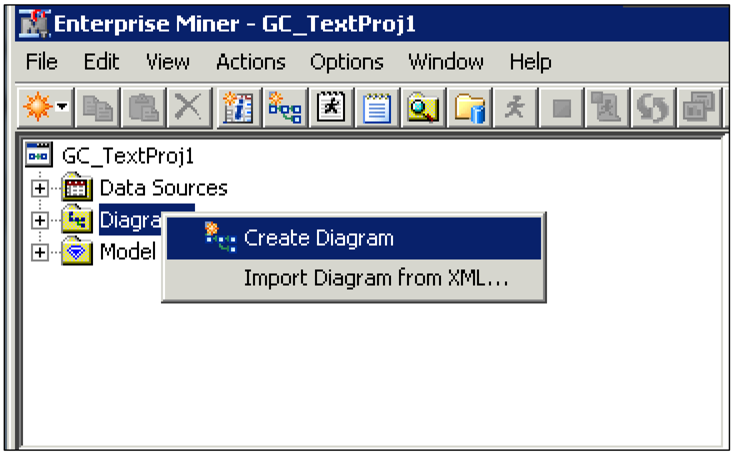
Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name: GC_Textproj1_Mar16

SAS Server Directory: C:\Temp

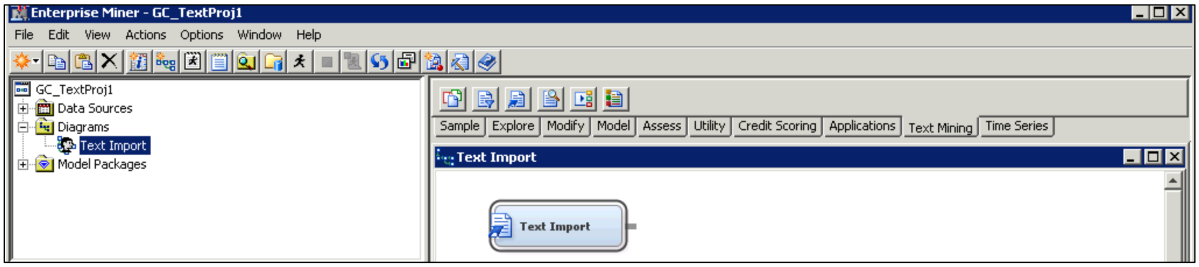
2. Click **Next** ⇨ **Next** to finish the project creation.

3. Right-click **Diagrams** in the project panel and select **New Diagram**. Give it a suitable name. Click **OK**.



Enterprise Miner - GC_TextProj1

4. Click and drag a **Text Import** node (from the Text Mining tab) onto the diagram workspace. Your diagram should look like this:



5. Click the ellipsis button next to **Import File Directory** in the properties panel of the Text Import node.

6. Point to the folder **Source_Files** that contains a number of SAS Global Forum papers in PDF format, some abstracts in Word format and a Powerpoint file etc.

7. Note the destination directory path. This folder will have the text files (created from PDF files by the File Import node and SAS Document Conversion server).

8. Change the text size in the properties panel of the File Import node to **32000** to extract the maximum amount of text.

   Note: You need to change the text size only if the variable to be used by the Text Parsing node is the Text variable. If you are going to use the filtered variable, then you can leave this property at its default level.

9. Right-click the **Text Import** node and select **Run**. Examine the results.

   The Results window of the Import node shows various distribution plots that report the number of documents that were omitted or truncated, document languages, types, sizes, and a time series plot of when the document was created, modified, and accessed. These plots give a basic idea of how successful the import task was. The Output window (reports the PROC FREQ results on document languages and a crosstabulation of omitted versus truncated documents. You need to pay attention to this window to figue out if any file was truncated (2) or omitted (none) and then explore the possible reasons for such truncation/omission.

10. Use Windows Explorer to take a quick look at how the files appear in the source files folder versus how the imported files appear in the destination folder.

11. Click the ellipsis button next to **Exported Data** in the properties panel of the Text Import node.

12. Select **TRAIN** and click **Properties** in the Exported Data - Text Import window.

13. Click the **Variables** tab to look at the variables in the exported data table.

Usually, it is the text parsing node that is connected to the text import node in the process flow. There are two columns that can be used by the Text Parsing node for identifying the text: FILTERED and TEXT. In the Text Parsing node, when you set the Use property of the Text variable to **Yes** and the Use property of the Filtered variable to **No**, only the text that is contained in the Text variable will be used for parsing. Otherwise, the Filtered variable is always given the preference when both the variables have the Use property set to **Yes**. Go back and click Explore in Step 12 and see for yourself what the data looks like coming out of the text Import node.

Bringing Textual Data into Enterprise Miner: How to Import Web Pages

Analyzing the data available on the web has become a business imperative for almost any organization. A treasure trove of information about brands, customer preferences, product reviews, stock prices, movie reviews, and so on, can be found on the World Wide Web. This information from the Internet in combination with internal data can make your analysis more powerful and current. However, the challenging part of the task is to find a way to get this data into SAS. Now, let's see how we can get data directly from the web into SAS Enterprise Miner. For this example, we will use analytics.okstate.edu as a website.

1. Add another Text Import node to the diagram and name it as Web Import.
2. Using Windows Explorer, create a new folder in **yourdrive:\Temp** with the names **WebImport_Source**
3. Change the node properties of the Text Import node for web crawling as shown below.
4. Set the import file directory to **C:\Temp\WebImport1_Source**.
5. Set the text size to **1000**. (Again, if you are going to analyze the text variable instead of the filtered variable, then this is not needed.)
6. Set the URL to **business.okstate.edu/analytics/**

   In a way similar to importing PDF files, as seen in the earlier demonstration, we would require two folders: a source folder for storing the downloaded html pages and the destination folder for holding the textual extracts. The path to the two folders that we created on our machine is set as the property value for the import file directory and destination file directory settings as shown above. Setting the Text Size property to 1000 will show a maximum text of 1,000 characters for each extracted web page from the website mentioned under URL. A value of 1 for the Depth property would extract all the files linked to from the starting web page only (analystics.okstate.edu). A value of Restricted in the domain will prevent the crawler from processing documents outside the domain of the initial web page.
7. Right-click and run the node.
8. Examine the results.

   You should also explore the data using the Exported Data ellipsis button from the properties panel to get a sense of what will be passed on to subsequent nodes in the diagram via the exported SAS data from this node.

End of Demonstration