

Step-by-Step SAS Text Miner Instructions for Text Clustering

We use SAS Text Miner to conduct text clustering. Text Miner is a component of SAS Enterprise Miner. This material is based on Enterprise Miner 14.1.

- Practice Data: Movie Text Reviews (45MB; Refer to the data description to understand the data.)

Download this dataset into your data directory folder.

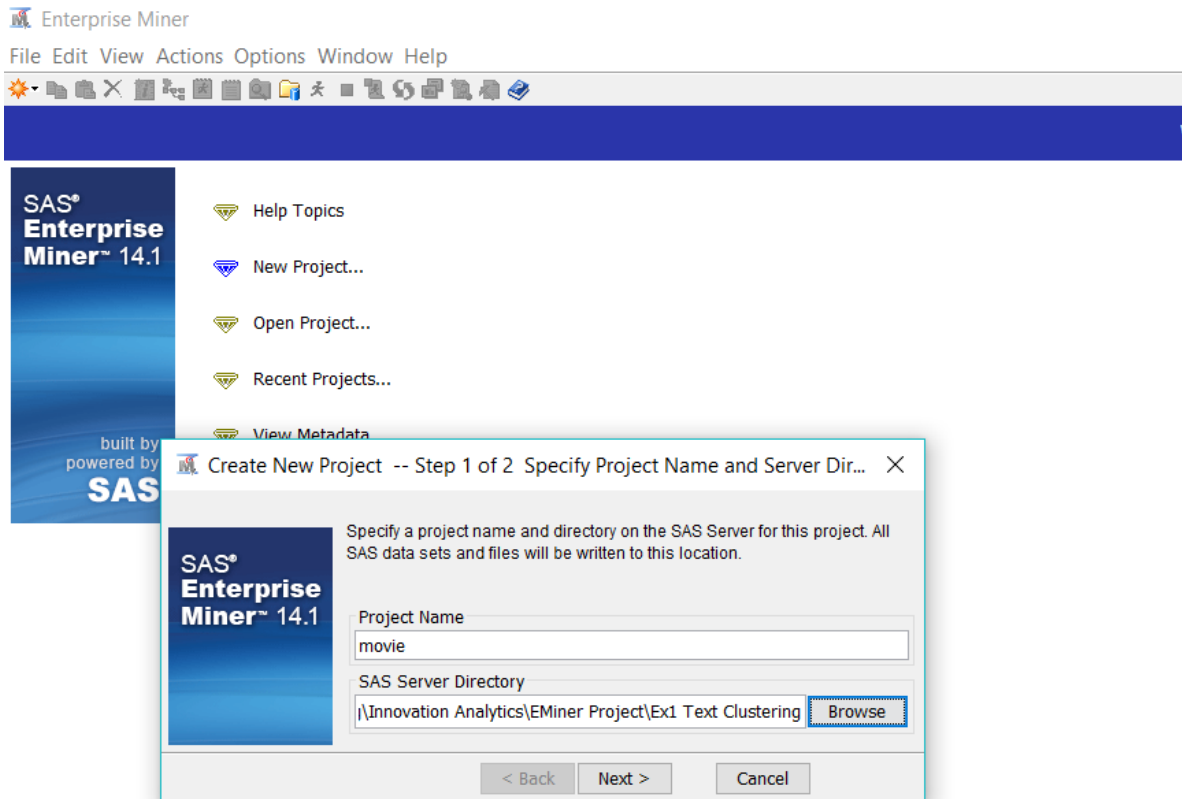
To open SAS Text Miner, choose the SAS program named *Enterprise Miner*.



Choose *New Project*.

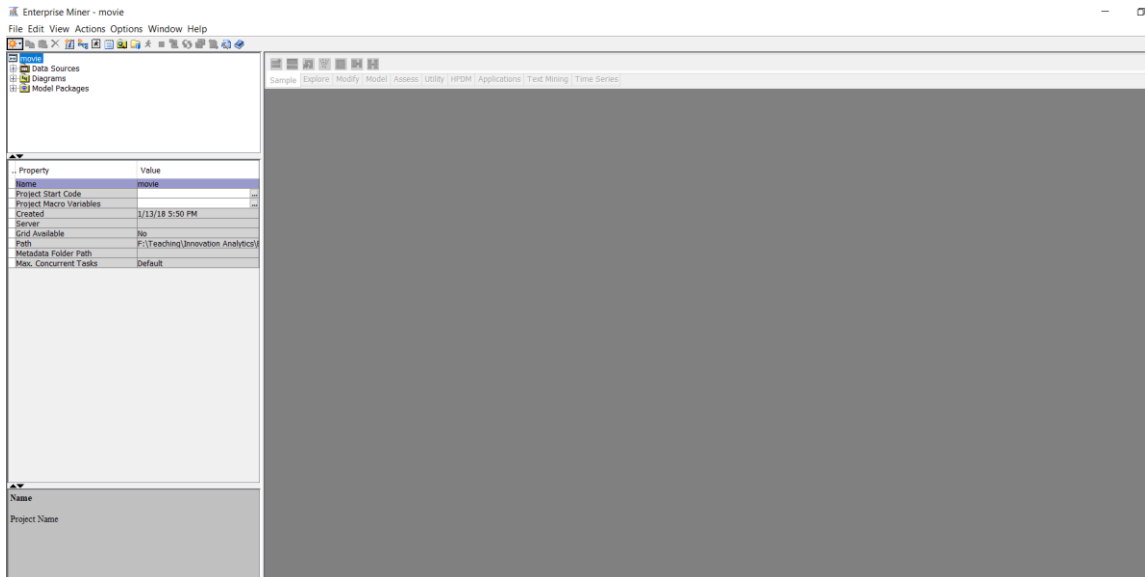
Project Name: type in “movie”

SAS Server Directory – Type in the address of your target directory folder, which will store all the results files from SAS once the whole procedure is completed.

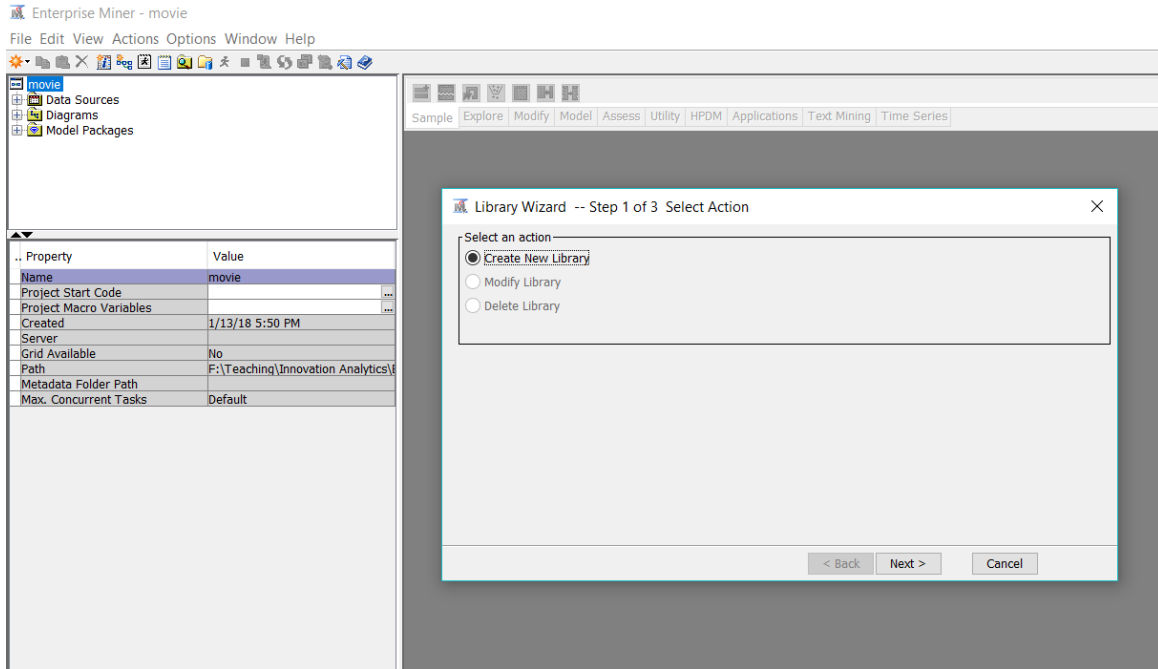


Choose *Next > Finish.*

Now, you will see a workspace named “Enterprise Miner – movie.”



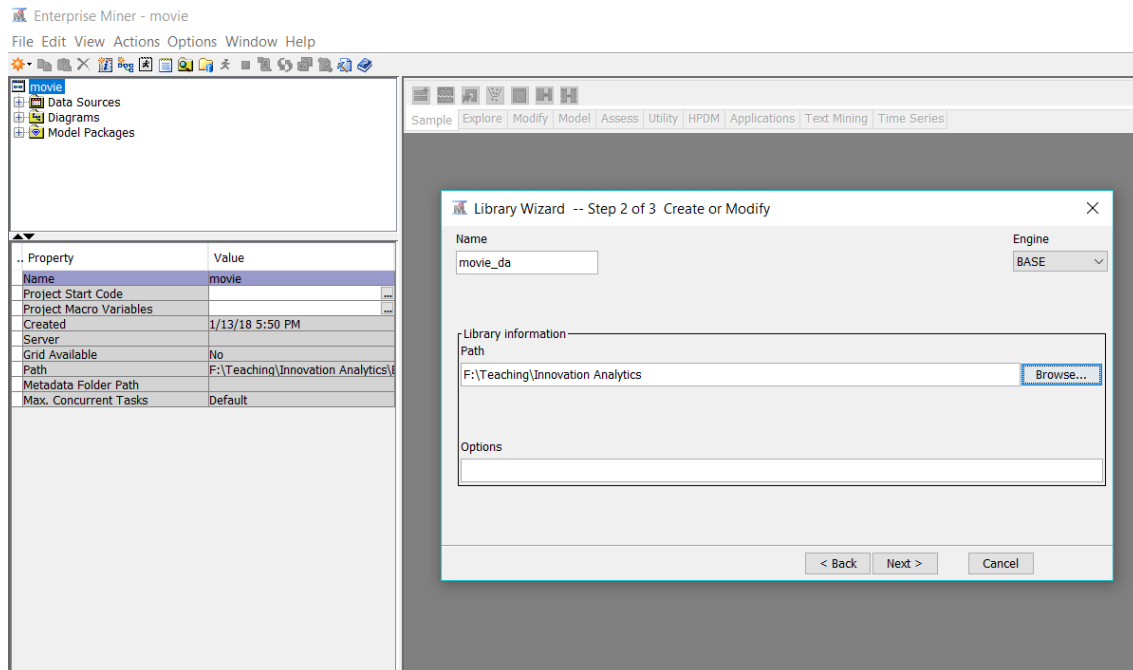
To set up our data access,
File < New < Library < Create New Library



Library Wizard

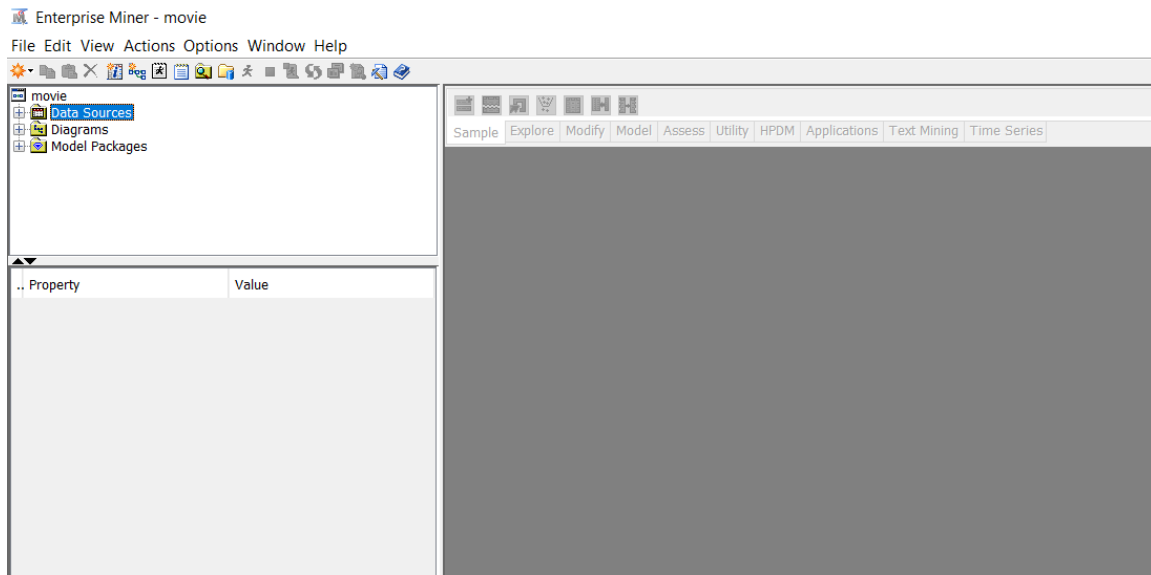
Name: movie_da

Path: Type in (your data directory folder address).

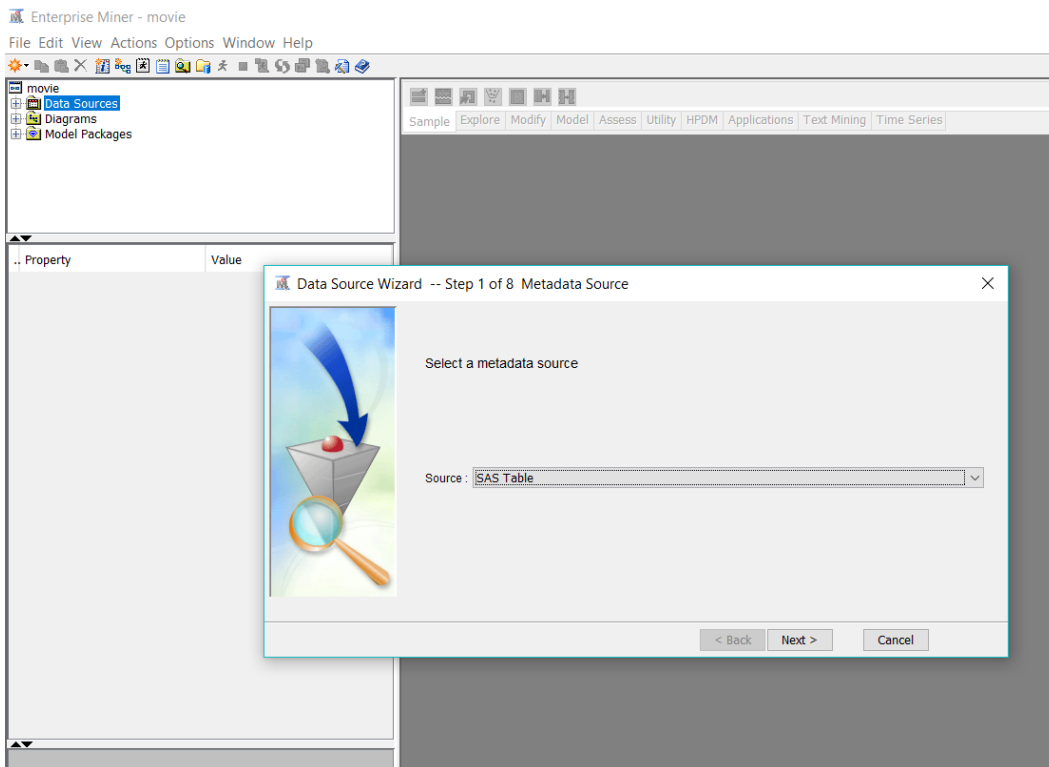
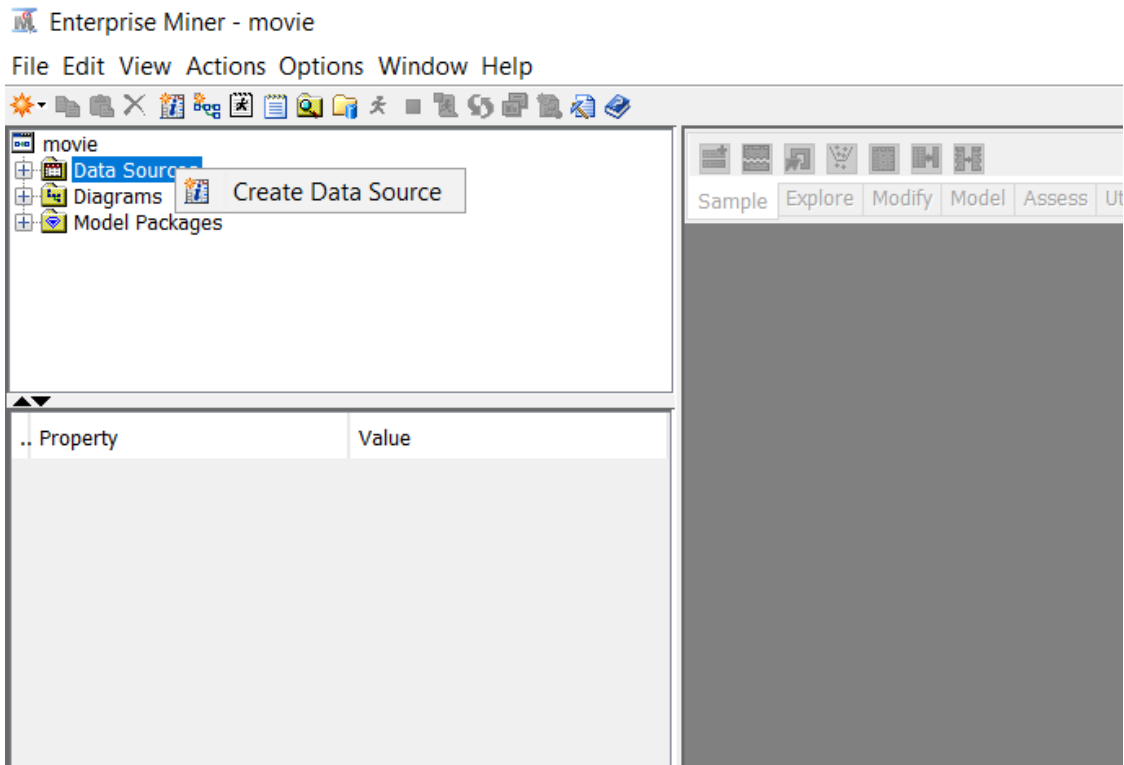


Hit *Next > Finish*.

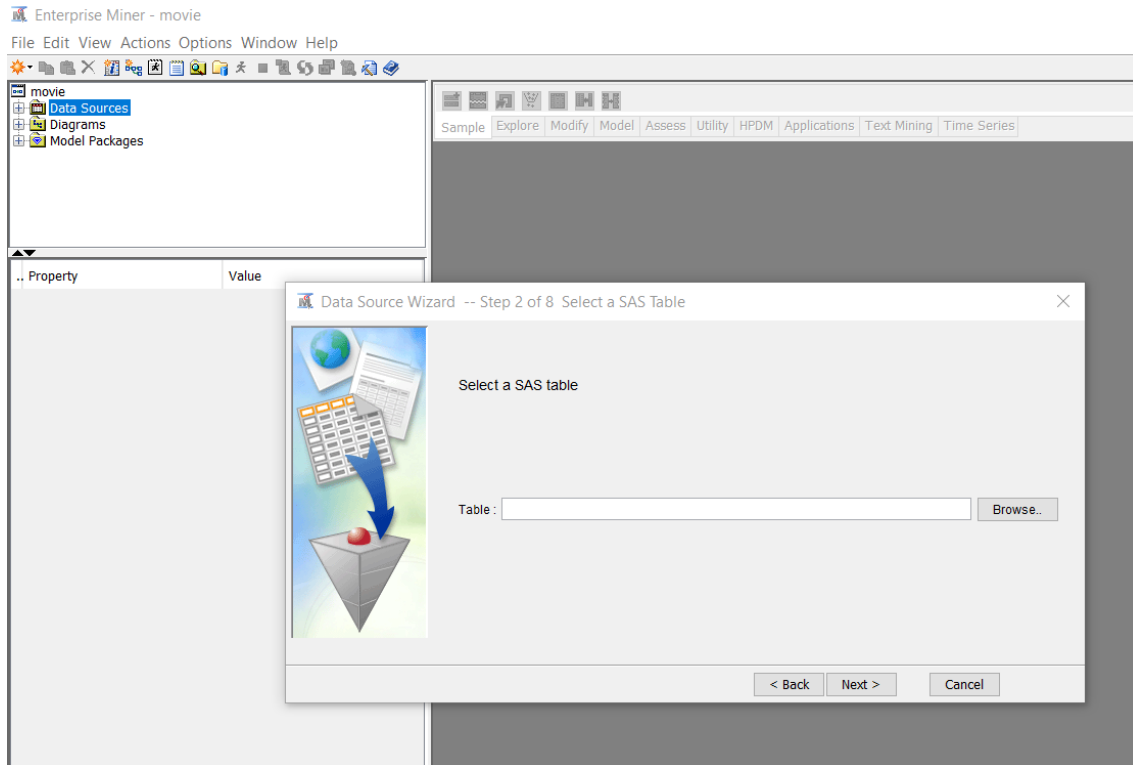
Place the mouse pointer on *Data Sources*.



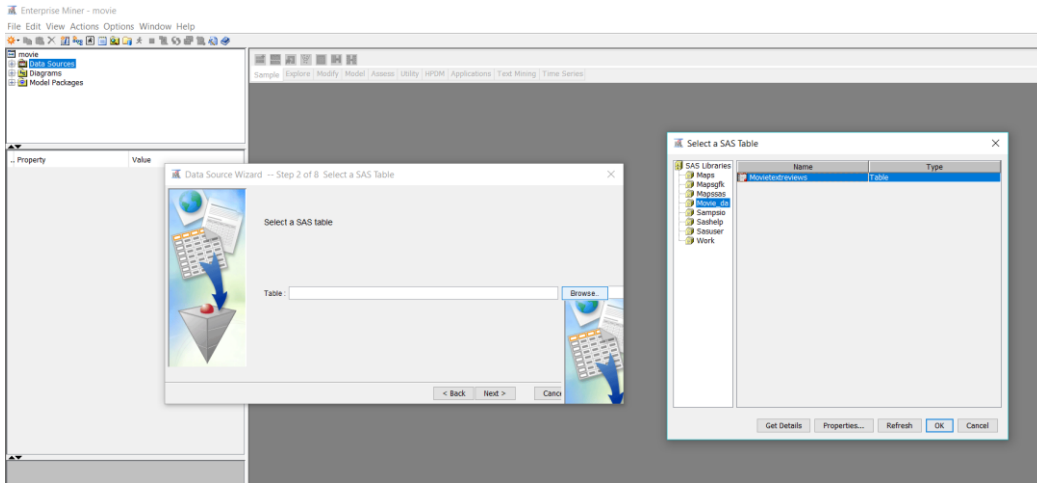
Find *Create Data Source* (with right click on *Data Sources*).



Hit Next.



Choose the target dataset in the created library.



Hit *OK* > *Next*.

For the rest of the steps in the wizard, choose the default options. Then, you can see the target dataset in the *Data Sources* section.

The screenshot shows the Enterprise Miner interface. The title bar reads "Enterprise Miner - movie". The menu bar includes "File", "Edit", "View", "Actions", "Options", "Window", and "Help". The toolbar contains various icons for file operations and analysis. The left pane shows a tree view with "Data Sources" expanded to "MOVIE_TEXTREVIEWS". The right pane shows a table of properties for the selected dataset.

Property	Value
ID	movietextreviews
Name	MOVIE_TEXTREVIEWS
Variables	...
Decisions	...
Role	Raw
Notes	...
Library	MOVIE_DA
Table	MOVIE_TEXTREVIEWS
Sample Data Set	
Size Type	
Sample Size	
Type	DATA
No. Obs	2799
No. Cols	10
No. Bytes	45859840
Segment	
Created By	Sangkil
Create Date	1/13/18 6:00 PM
Modified By	Sangkil
Modify Date	1/13/18 6:00 PM
Scope	Local

Now, to set up Text Miner,
Diagrams < *Create Diagram* (with right click on *Diagrams*)

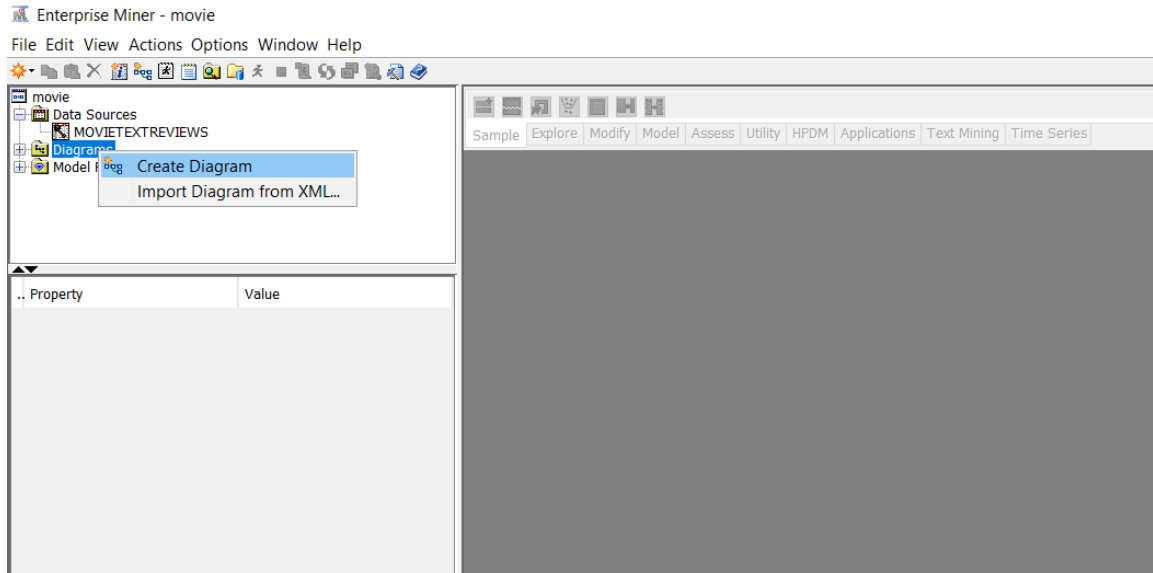
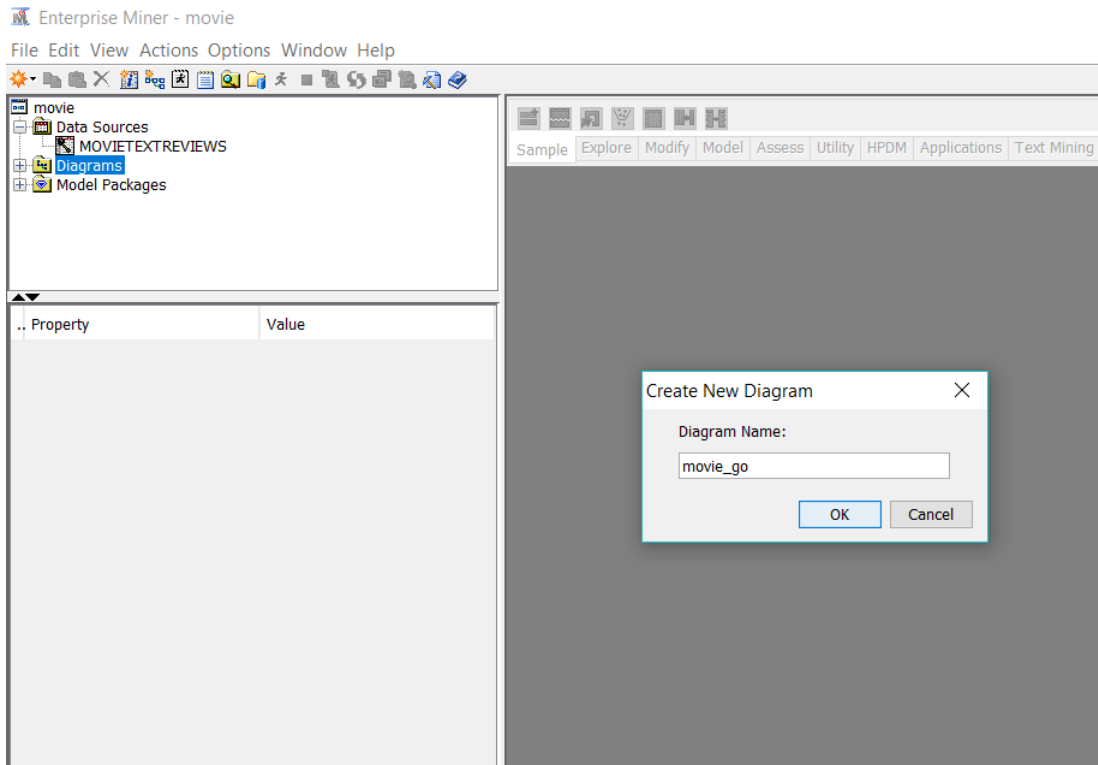
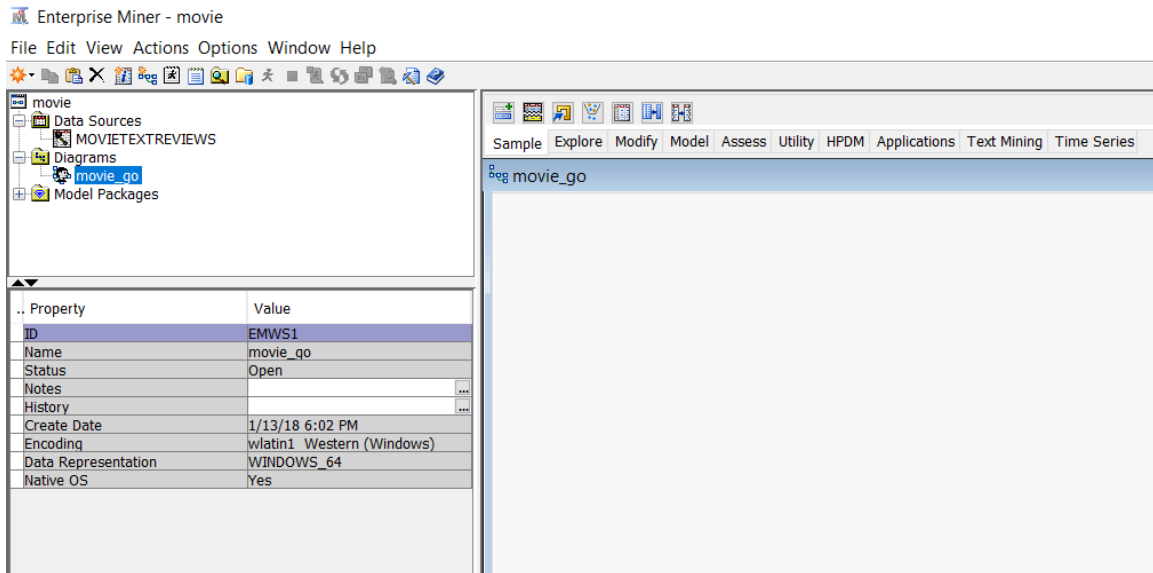


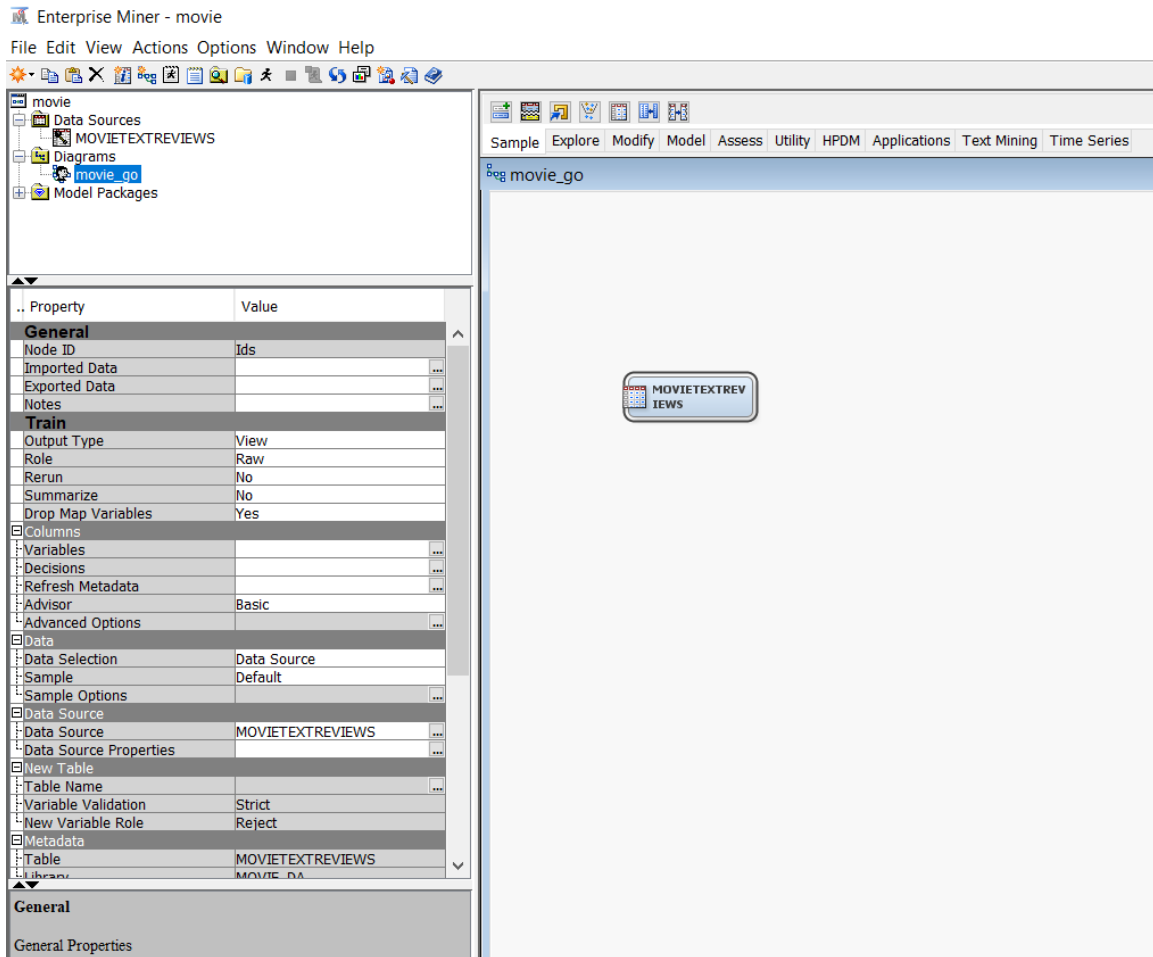
Diagram Name: Enter 'movie_go.'



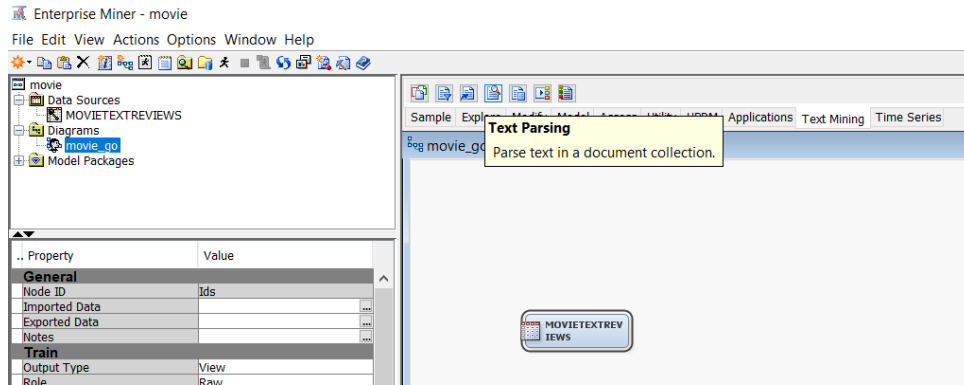
Hit *OK*. Then, you will see a white workspace named “movie_go.”



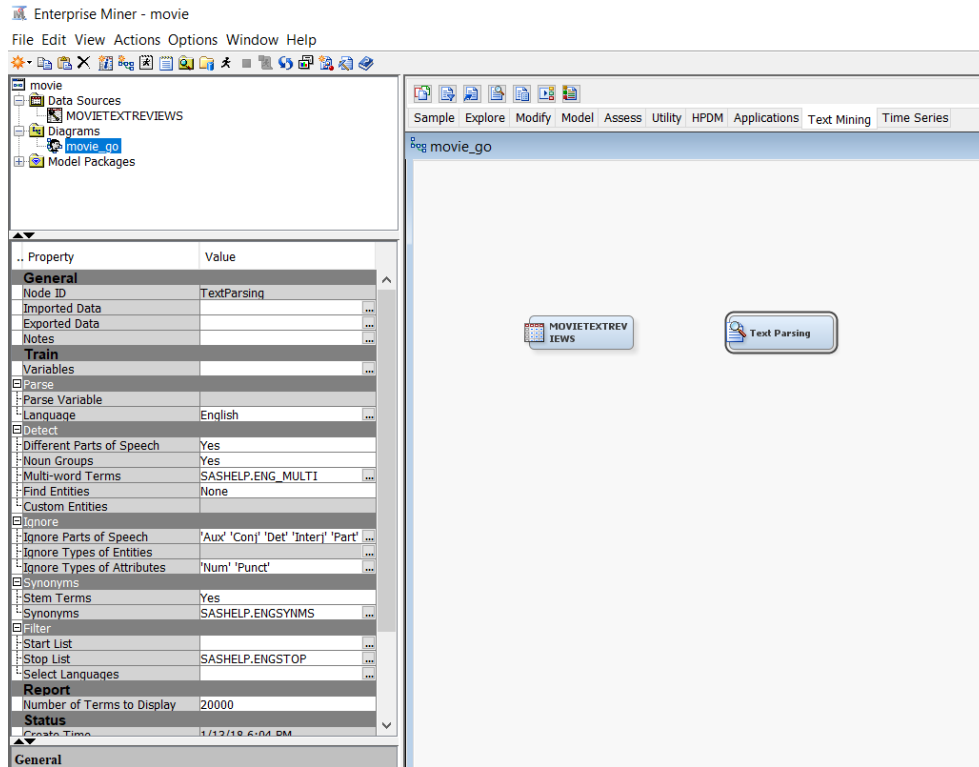
Drag the target data onto the workspace.



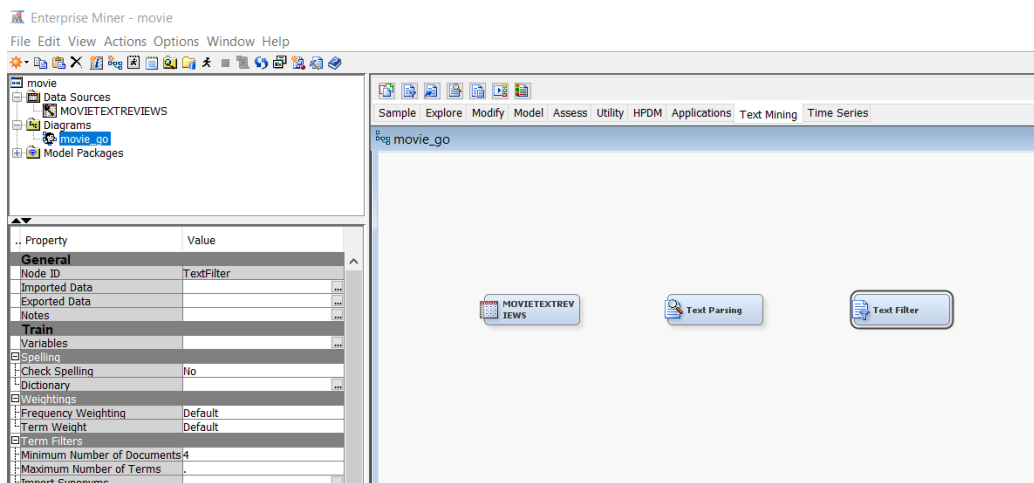
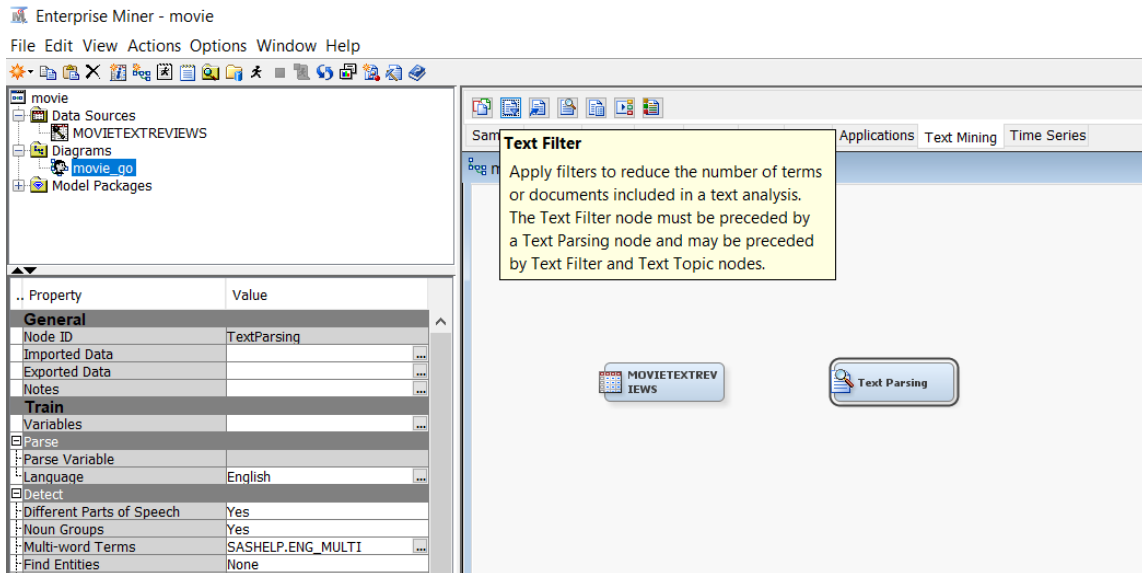
Choose the *Text Mining* tab in the top menu.



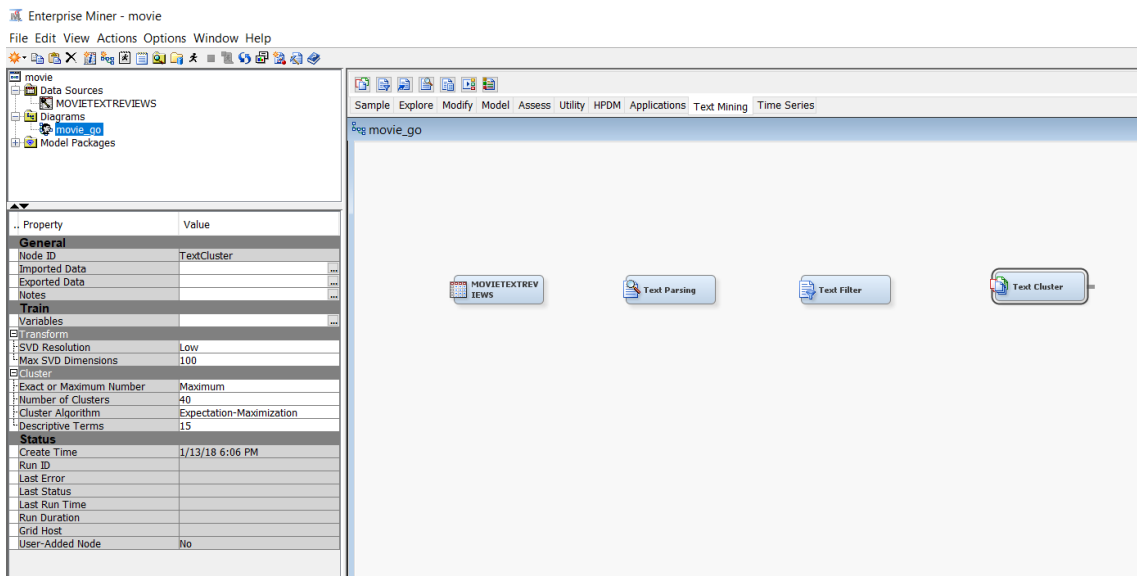
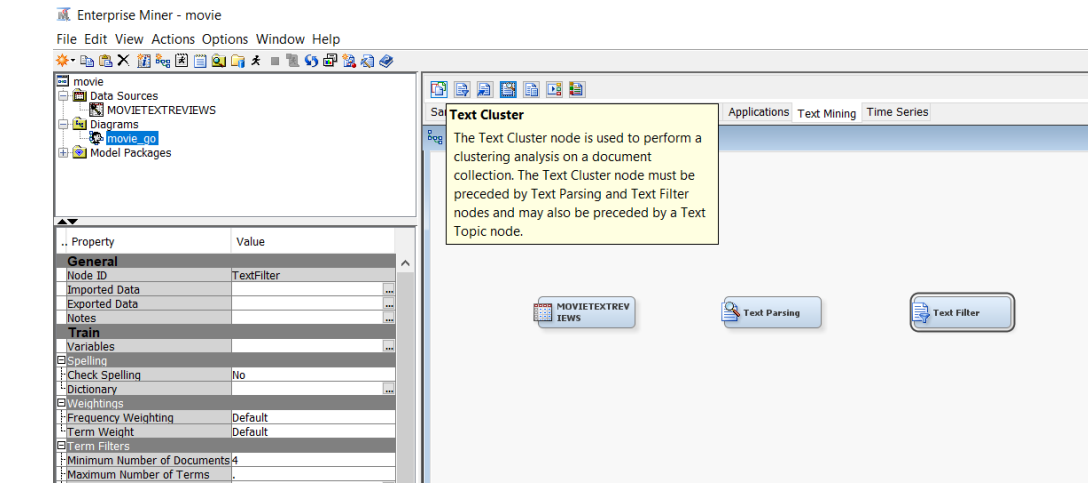
Choose the *Text Parsing* node and drag it onto the workspace.



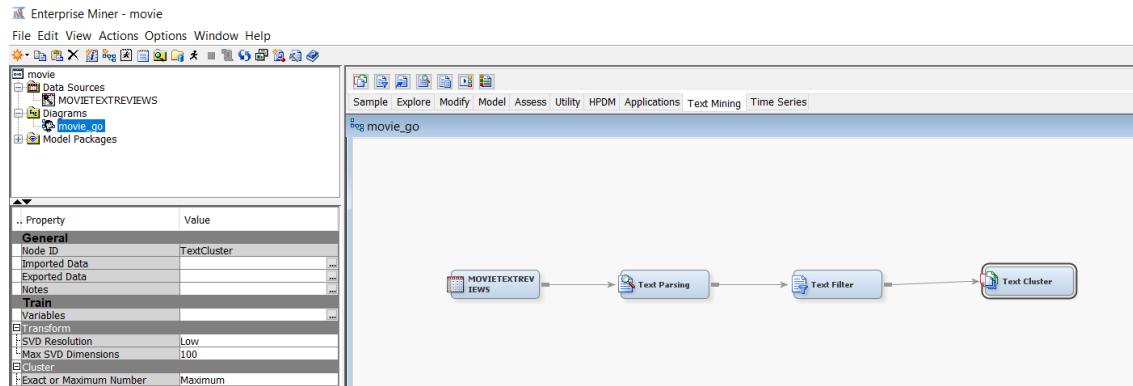
Choose the *Text Filter* node and drag it onto the workspace.



Choose the *Text Cluster* node and drag it onto the workspace.



Link the data to the *Text Parsing* node.
Link the *Text Parsing* node to the *Text Filter* node.
Link the *Text Filter* node to the *Text Cluster* node.



To run the required text clustering analysis,

Click the *Text Cluster* node on the workspace. Then, you will see the detailed operation information of the node on the lower left side.

Train < Cluster

Exact or Maximum Number: Choose 'Exact.'

Determine your *Number of Clusters*. (2 in our case)

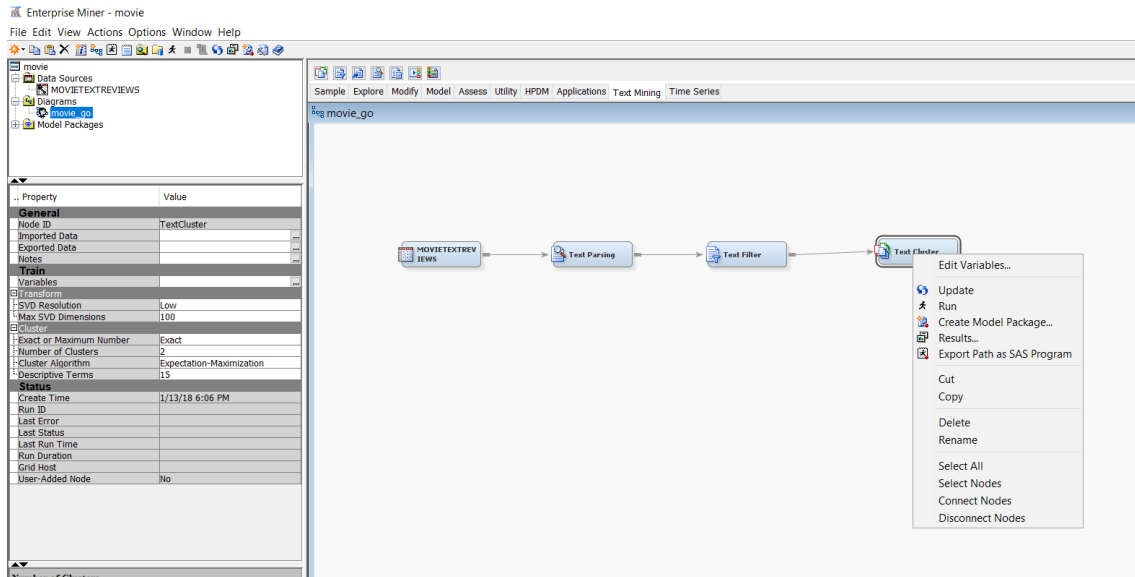
Determine your number of *Descriptive Terms*. (15 is the default option.)

The screenshot shows the Enterprise Miner interface. On the left, a tree view shows the project structure: 'movie' > 'Data Sources' > 'MOVIEEXTREVIEWS' > 'Diagrams' > 'movie-go'. Below this is a table with columns 'Property' and 'Value'.

Property	Value
General	
Node ID	TextCluster
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Techniques	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	2
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	1/13/18 6:06 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

The main workspace shows a workflow diagram with four nodes: 'MOVIEEXTREVIEWS', 'Text Parsing', 'Text Filter', and 'Text Cluster', connected by arrows from left to right. The 'Text Cluster' node is highlighted with a blue border.

After placing the mouse pointer on the Text Cluster nod, choose *Run* with right click.



It may take some time to generate the text clustering results (usually, a few minutes).

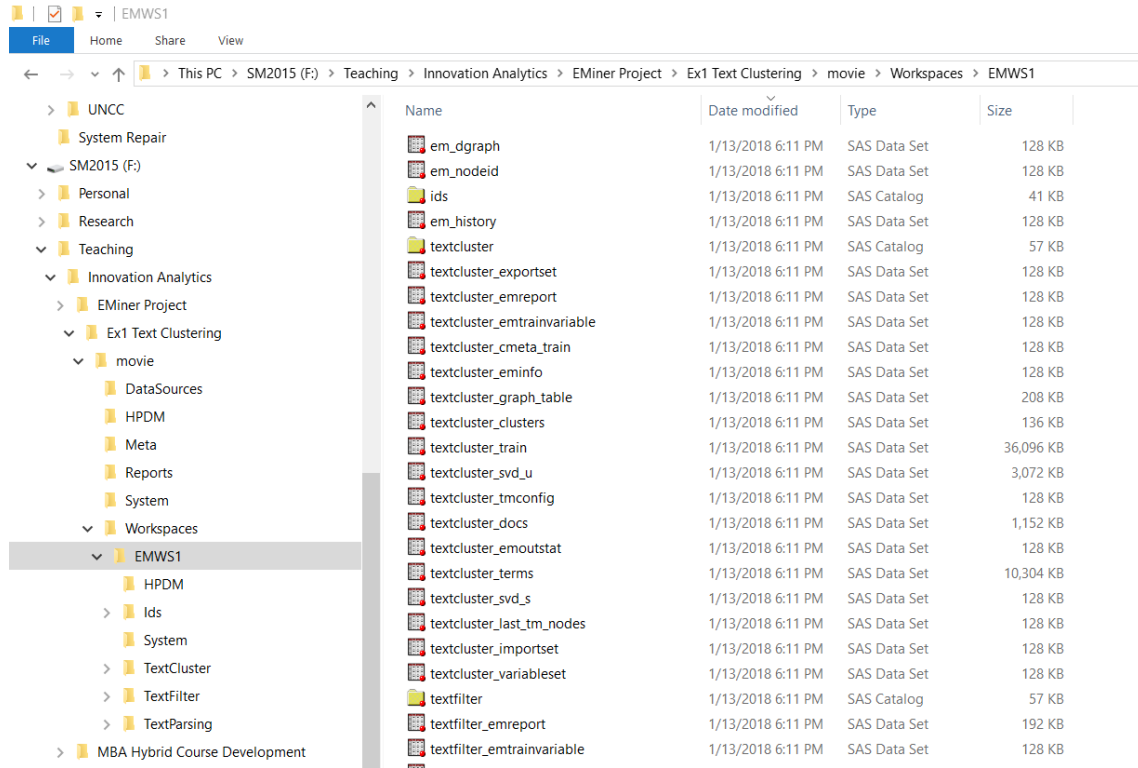
The screenshot displays the Enterprise Miner interface. On the left, a tree view shows the project structure with 'movie' selected. Below it, a property table is visible:

Property	Value
General	
Node ID	TextCluster
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	2
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	1/13/18 6:06 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

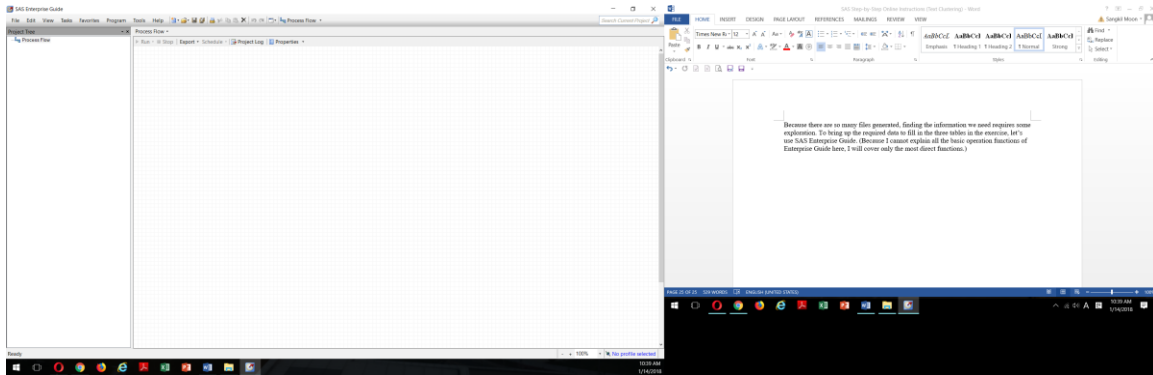
The main workspace shows a workflow diagram with four nodes: 'MOVIEEXTREVIEW REVIEWS', 'Text Parsing', 'Text Filter', and 'Text Cluster'. A 'Confirmation' dialog box is open, asking: 'Do you want to run this path? Diagram: movie_go Path: Text Cluster'. The dialog has 'Yes' and 'No' buttons.

Hit *Yes* to see some results. The results you see here is only a small fraction of the results the program have generated, which can be found in the following directory address:

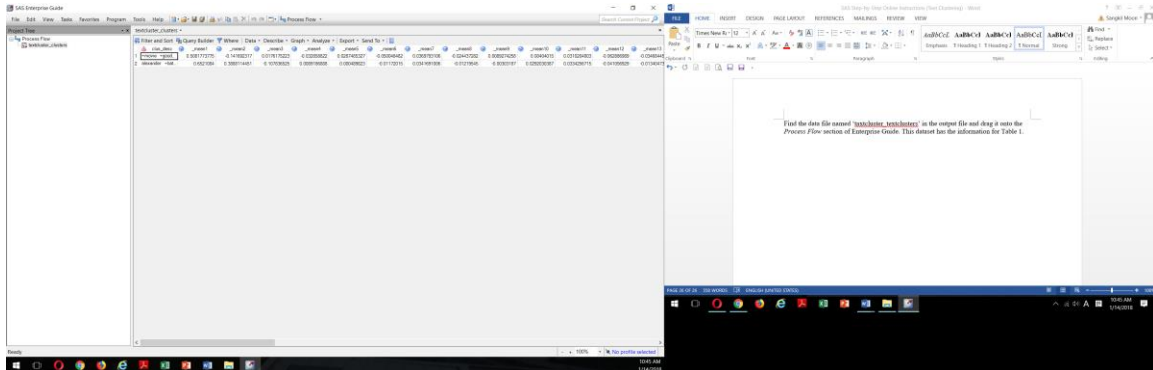
(Your project target directory folder address) < movie < Workspaces < EMWS1



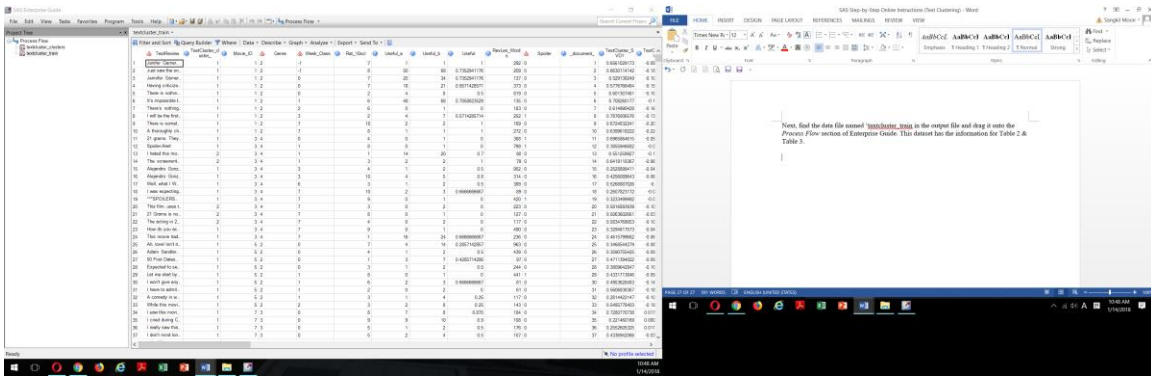
Because there are so many files generated, finding the information we need requires some exploration. To bring up the required data to fill in the three tables in the exercise, let's use *SAS Enterprise Guide*. (Because I cannot explain all the basic operation functions of Enterprise Guide here, I will cover only the most direct functions.)



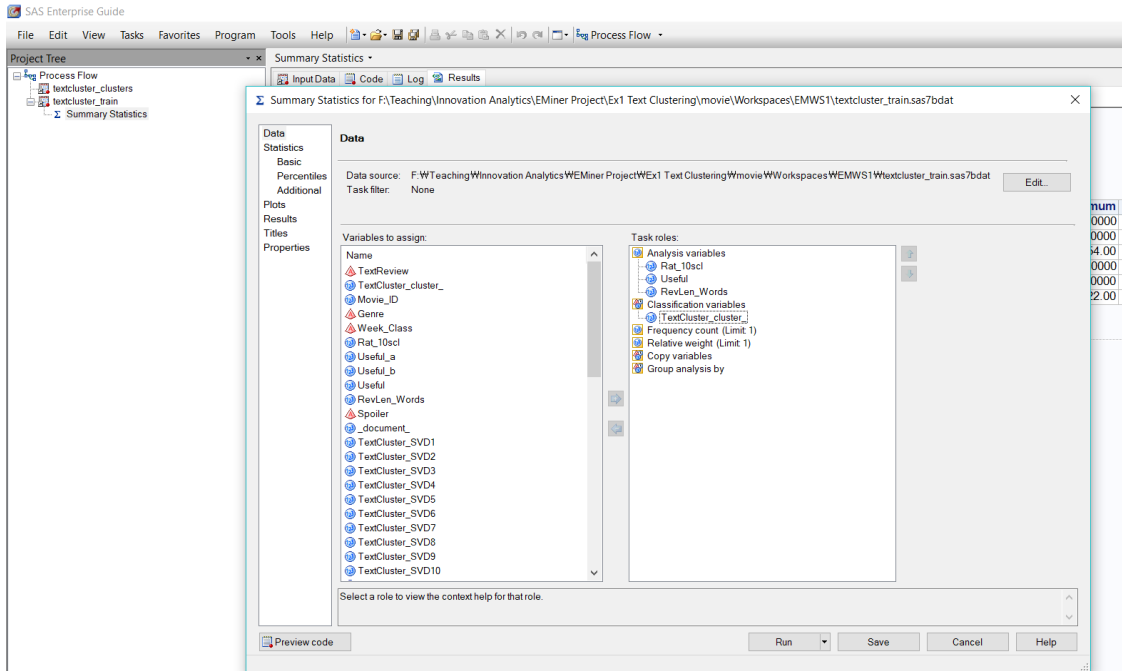
Find the data file named 'taxtcluster_textclusters' in the output file and drag it onto the *Process Flow* section of Enterprise Guide. This dataset has the information for Table 1.



Next, find the data file named 'taxtcluster_train' in the output file and drag it onto the *Process Flow* section of Enterprise Guide. This dataset has the information for Table 2 & Table 3.

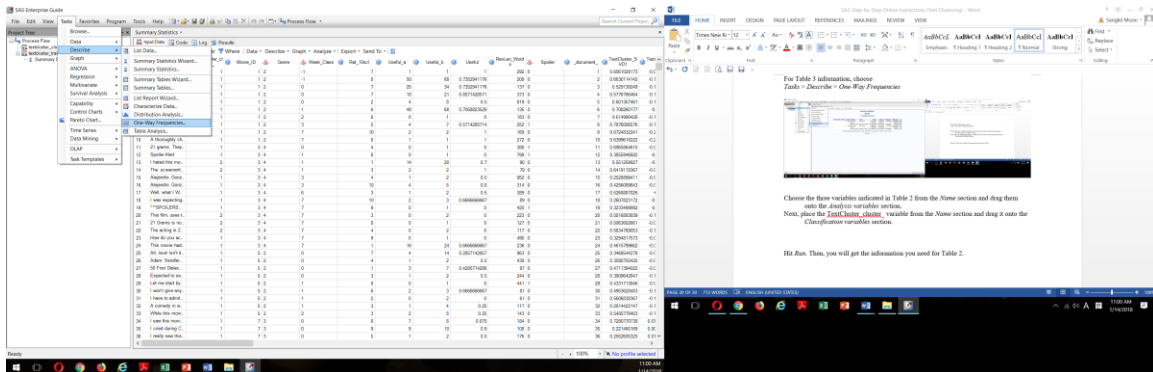


Choose the three variables indicated in Table 2 from the *Name* section and drag them onto the *Analysis variables* section.
Next, place the *TextCluster_cluster_* variable from the *Name* section and drag it onto the *Classification variables* section.

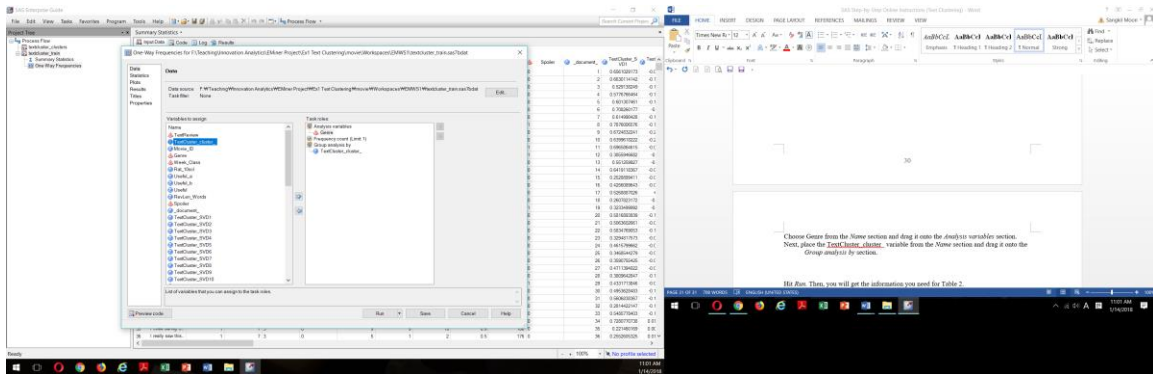


Hit *Run*. Then, you will get the information you need for Table 2.

For Table 3 information, choose
Tasks > Describe > One-Way Frequencies



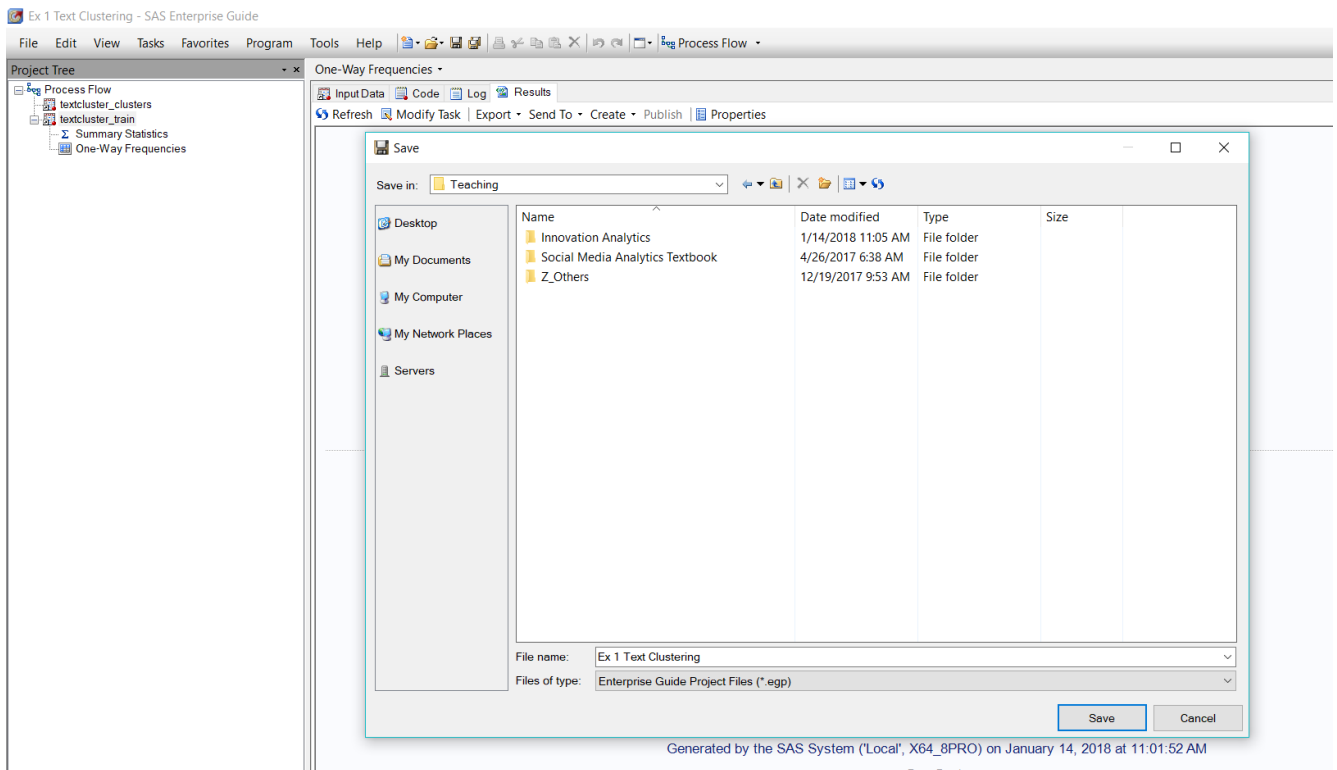
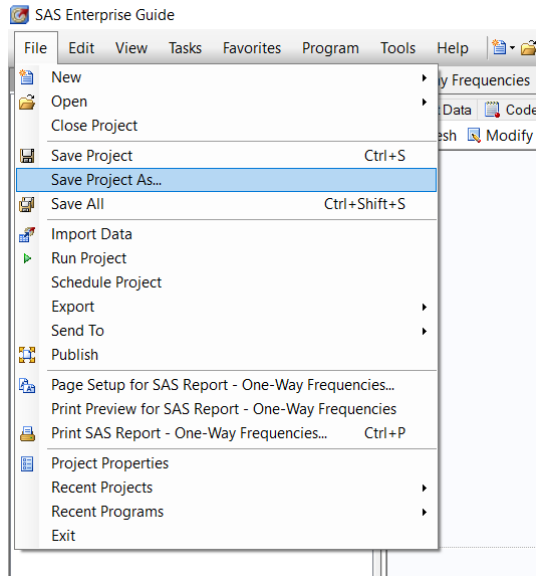
Choose Genre from the *Name* section and drag it onto the *Analysis variables* section. Next, place the TextCluster_cluster_variable from the *Name* section and drag it onto the *Group analysis by* section.



Hit *Run*. Then, you will get the information you need for Table 3.

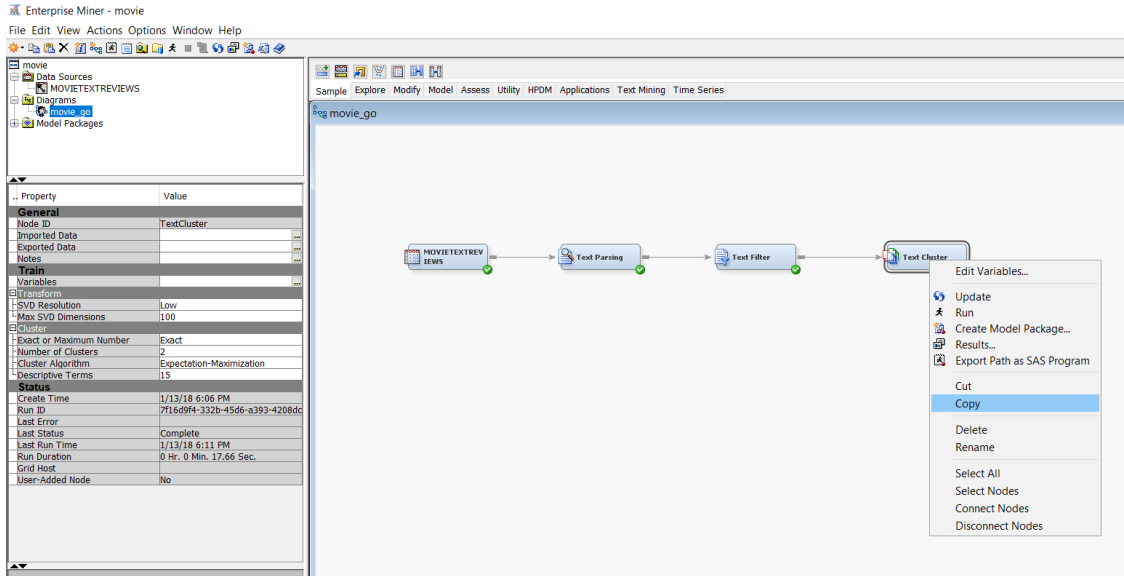
Save the whole content of your work on Enterprise Guide in the name of “Ex 1 Text Clustering” as a project by choosing

File > Save Project As

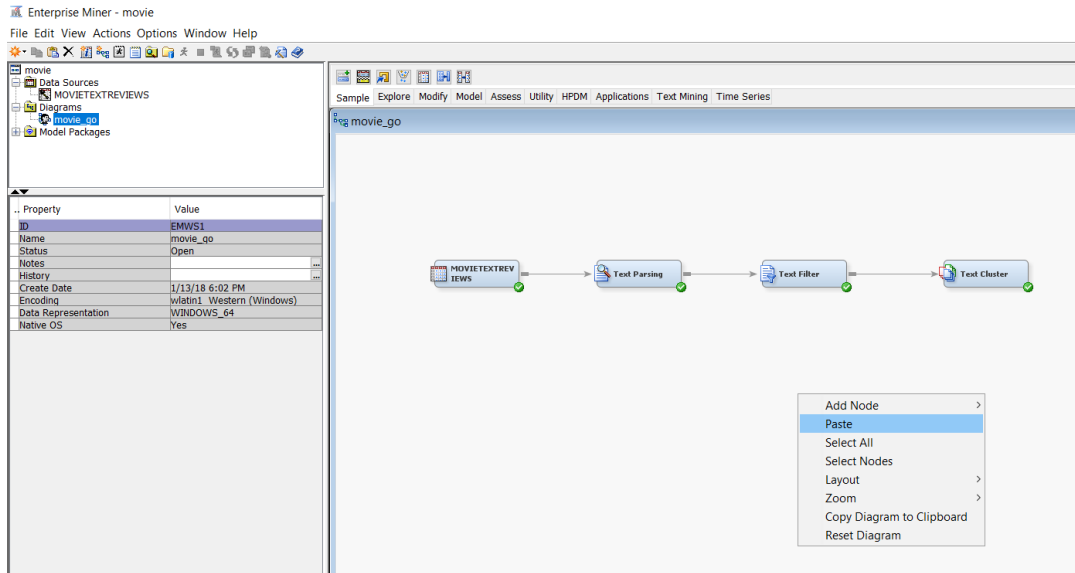


Now, you have obtained all the analysis results for **Task A: Text Clustering – 2 Clusters** in the exercise. You should use necessary information to actually complete Task A.

Next, to do **Task B: Text Clustering – 3 Clusters**, go back to the workspace named *movie_go* on *Enterprise Miner*.
Point to the *Text Cluster* node and choose *Copy* with right click.



Move the mouse pointer out of the node and place it at any empty space spot and choose *Paste* with right click.



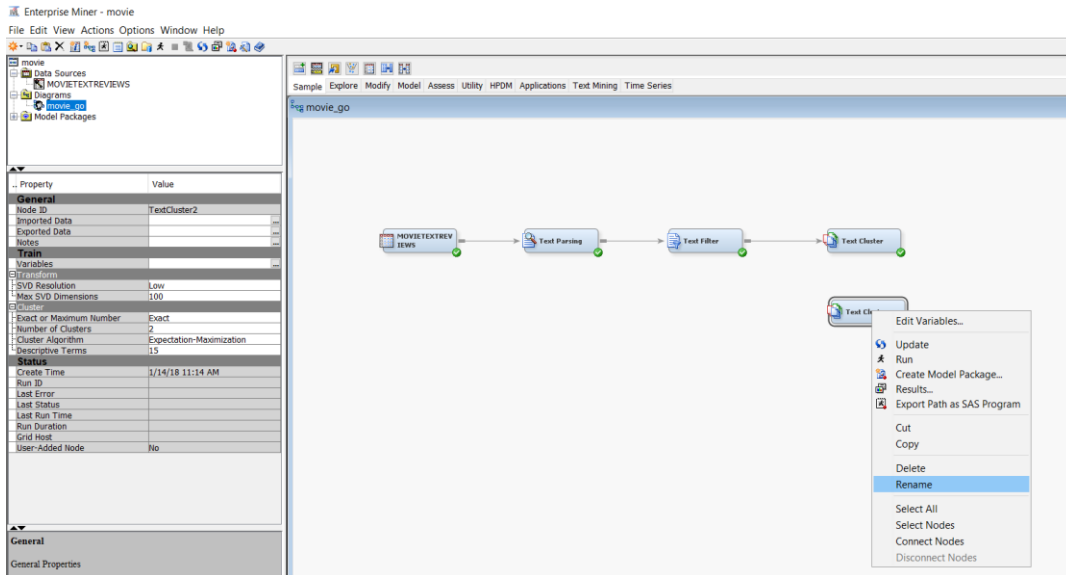
Then, you will get a new *Text Cluster* node.

The screenshot displays the Enterprise Miner interface. On the left, a tree view shows the project structure with 'movie' as the root, containing 'Data Sources' (MOVIETEXTREVIEWS), 'Diagrams' (movie.go), and 'Model Packages'. Below this is a properties table for the selected 'Text Cluster' node.

Property	Value
General	
Node ID	TextCluster2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	2
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	1/14/18 11:14 AM

The main workspace shows a workflow diagram titled 'movie.go'. The workflow consists of four nodes connected by arrows: 'MOVIETEXTREVIEWS' (input), 'Text Parsing', 'Text Filter', and 'Text Cluster'. A separate 'Text Cluster' node is also visible below the main flow.

Place the pointer on the new node and choose *Rename* with right click.



Type in “Text Cluster (3 clusters)” and hit *OK*.

The screenshot shows the Enterprise Miner interface. On the left, a tree view shows the project structure: 'movie' containing 'Data Sources' (with 'MOVIETEXTREVIEWS'), 'Diagrams', 'movie_go', and 'Model Packages'. Below the tree is a properties table for a 'Text Cluster' node.

Property	Value
General	
Node ID	TextCluster2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	2
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	1/14/18 11:14 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

The main workspace shows a workflow diagram with the following nodes: 'MOVIETEXTREVIEWS' (Data Source), 'Text Parsing' (Text Mining), 'Text Filter' (Text Mining), and 'Text Cluster' (Text Mining). A 'Text Cluster (3 clusters)' node is also visible in the workspace. The workflow is titled 'movie_go'.

Change *Number of Clusters* from 2 to 3 (at left).

The screenshot shows the Enterprise Miner interface. On the left, a tree view shows the project structure: 'movie' > 'Data Sources' > 'MOVIETEXTREVIEWS' > 'Diagrams' > 'movie_go'. Below this is a property table for the 'Text Cluster' node.

Property	Value
General	
Node ID	TextCluster2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	3
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	1/14/18 11:14 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

The main workspace shows a workflow diagram for 'movie_go'. The workflow consists of four nodes: 'MOVIETEXTREVIEWS', 'Text Parsing', 'Text Filter', and 'Text Cluster'. Each node has a green checkmark indicating it is active. Below the main workflow, there is a separate node labeled 'Text Cluster (3 clusters)'.

Connect *Text Filter* to *Text Cluster (3 clusters)* with an arrow.

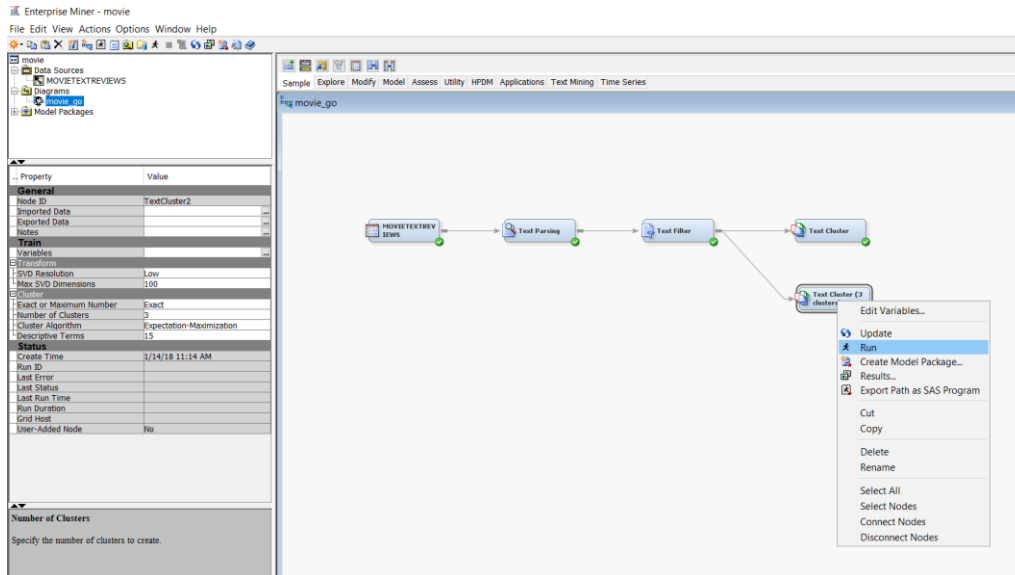
The screenshot shows the Enterprise Miner interface. On the left, a tree view shows the project structure: 'movie' > 'Data Sources' > 'MOVIEEXTREVIEWS' > 'Diagrams' > 'movie_go'. Below this is a 'Property' table for the selected 'TextCluster2' node.

Property	Value
General	
Node ID	TextCluster2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	3
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	1/14/18 11:14 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

The main workspace shows a workflow diagram with the following nodes and connections:

- 'MOVIEEXTREVIEWS' (Data Source) connects to 'Text Parsing' (Transform).
- 'Text Parsing' connects to 'Text Filter' (Transform).
- 'Text Filter' connects to 'Text Cluster' (Cluster).
- 'Text Filter' also connects to 'Text Cluster (3 clusters)' (Cluster).

Place the pointer on the new node and choose *Run* with right click.



The rest of the procedure should remain the same with the previous 2-cluster case. Lastly, in the same output folder,

(Your project target directory folder address) < movie < Workspaces < EMWS1

you will see the newly added data files with number 2 (with the second node operation this time) such as “textcluster2_clusters” and “textcluster2_train” this time.

