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.)

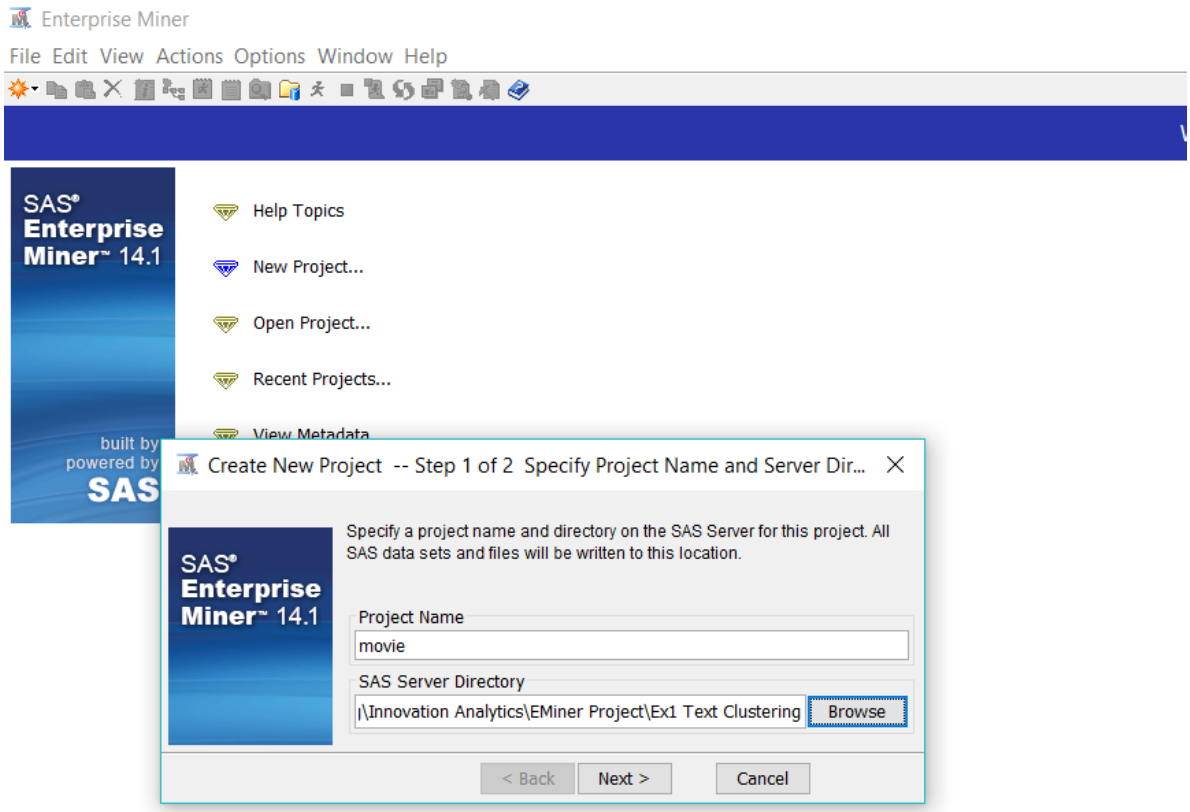
Download this dataset into your data directory folder.

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

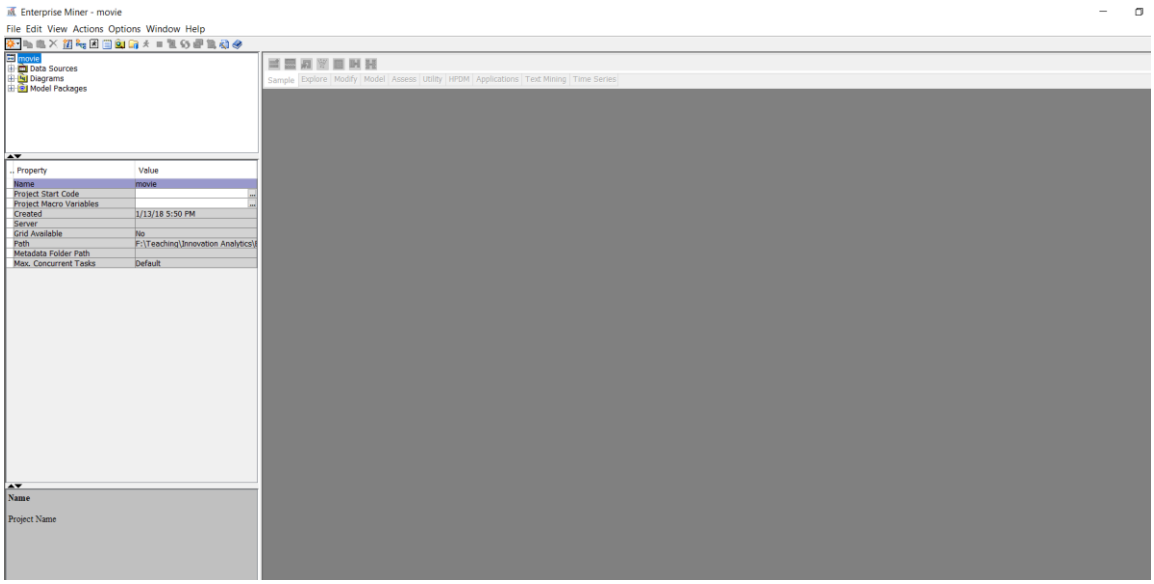
Choose *New Project*.

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.

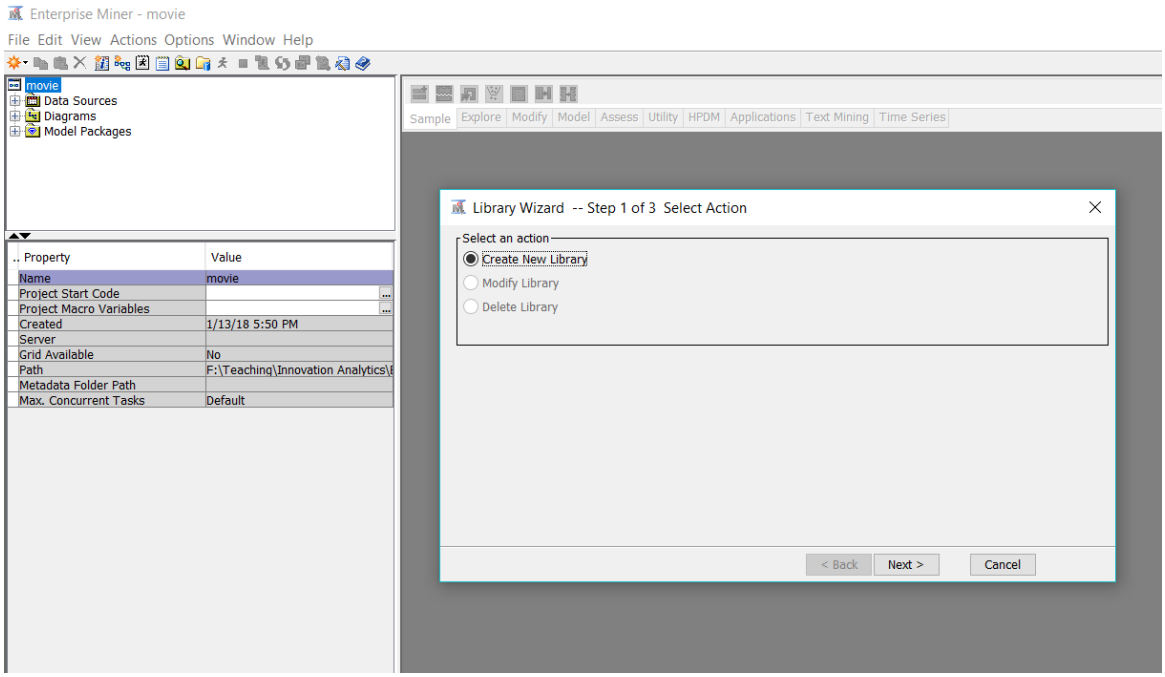


Choose *Next > Finish.*
Now, you will see a workspace named “Enterprise Miner – movie.”



To set up our data access,

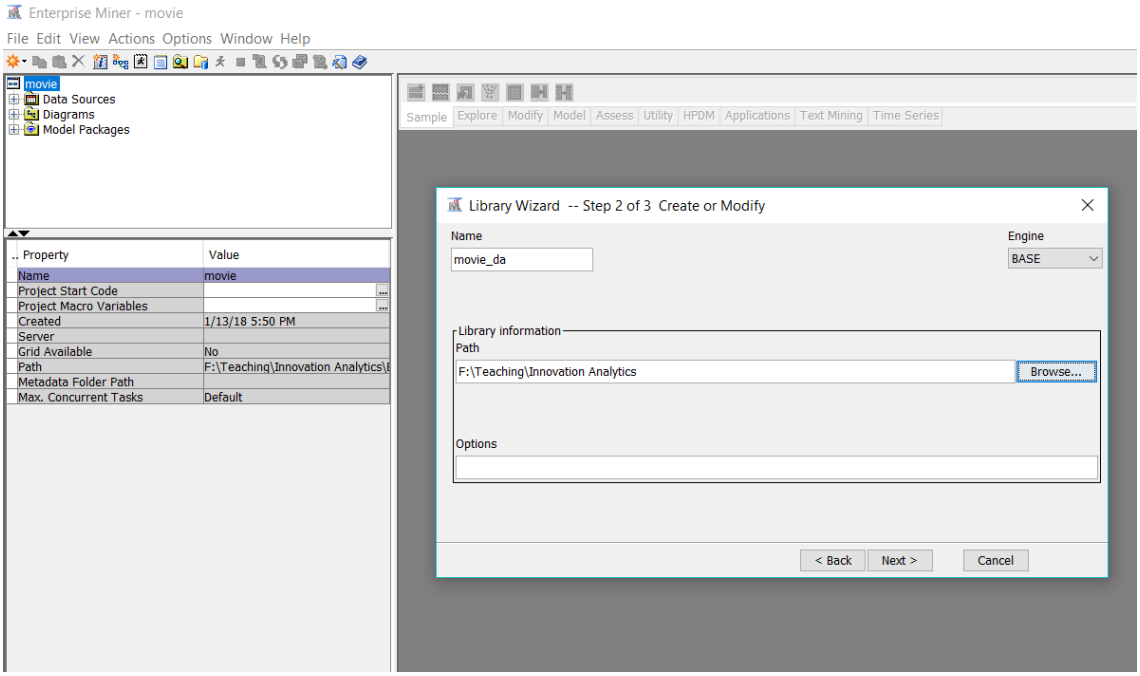
File < New < Library < Create New Library



Library Wizard

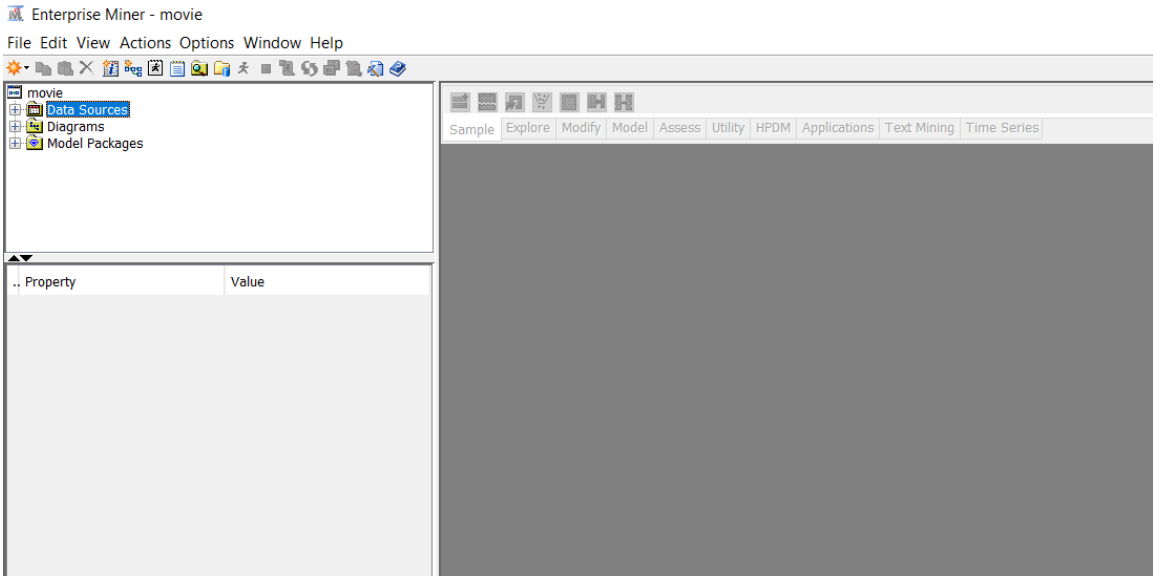
Name: movie_da

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

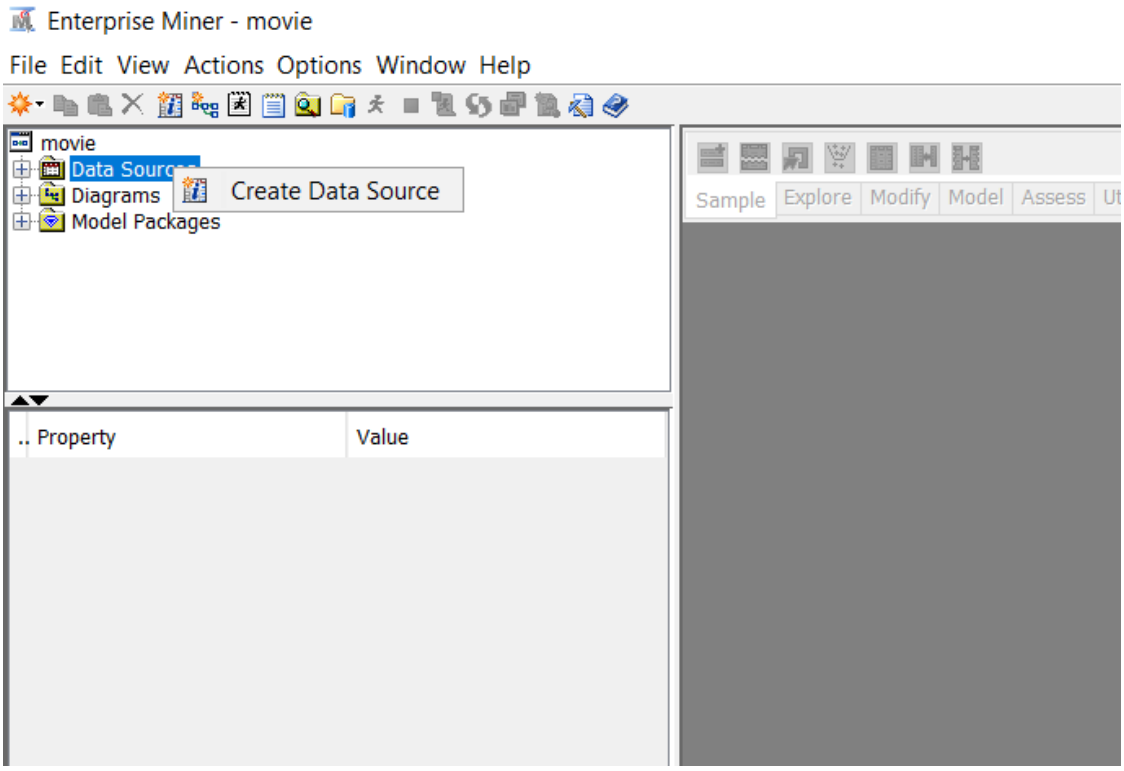


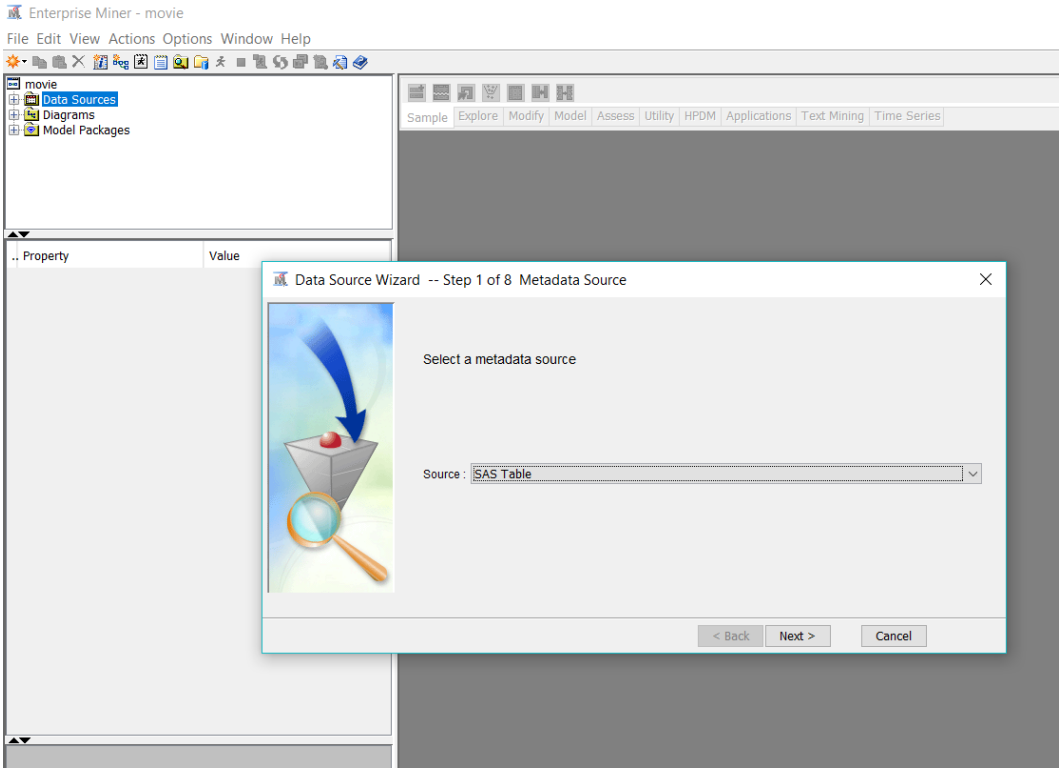
Hit *Next* > *Finish*.

Place the mouse pointer on *Data Sources.*

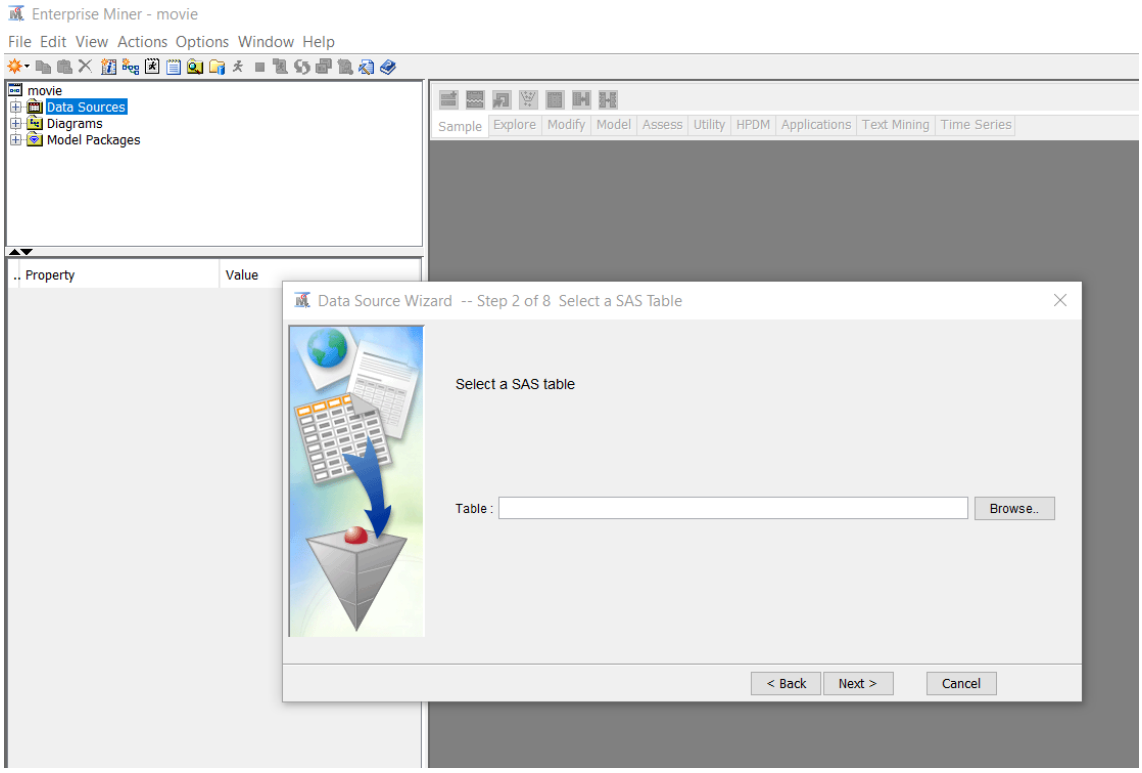


Find *Create Data Source* (with right click on *Data Sources*).

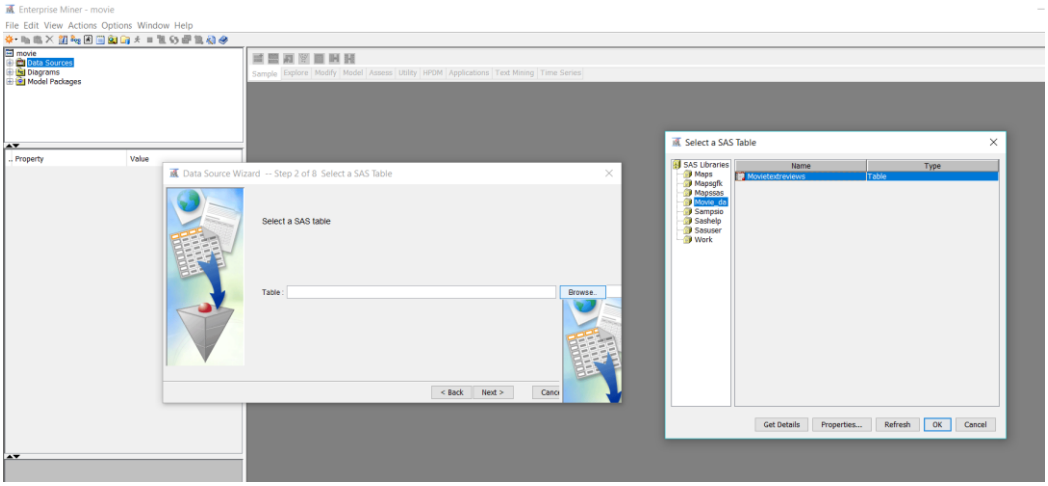




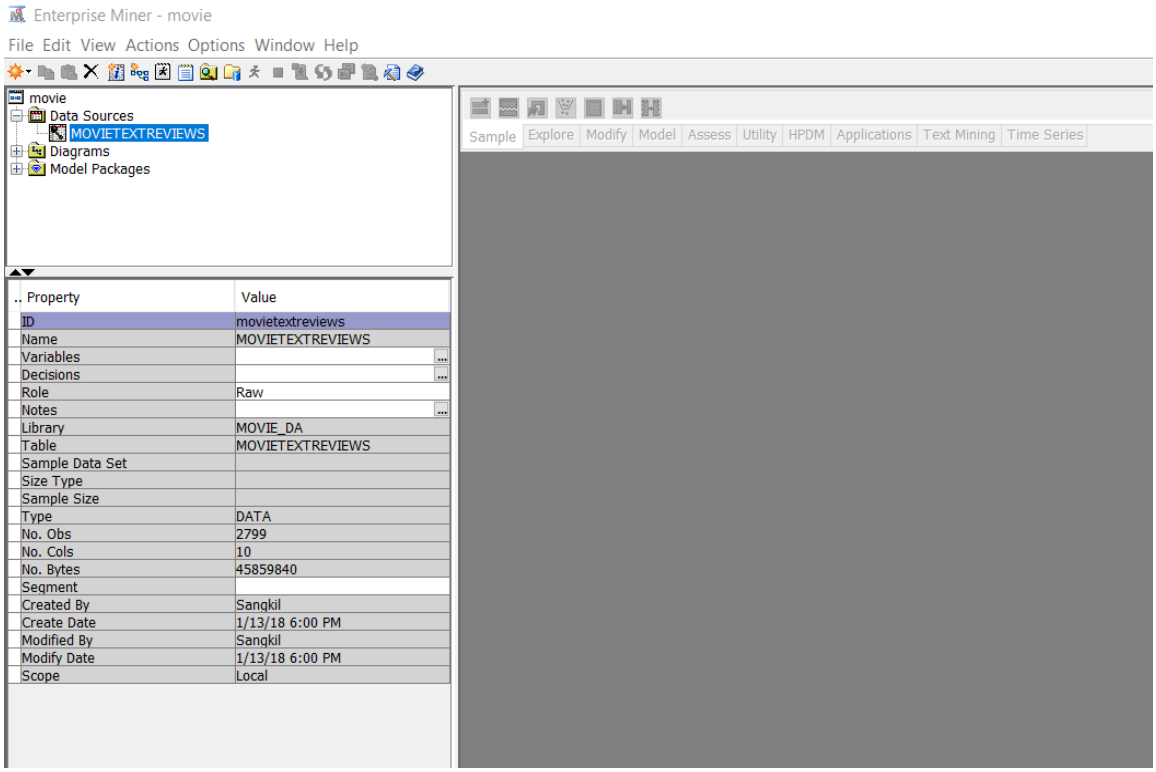
Hit *Next*.


Enterprise Miner - movie
File Edit View Actions Options Window Help
movie
Data Sources
Diagrams
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Property
Value
Data Source Wizard -- Step 2 of 8 Select a SAS Table
Select a SAS table
Table :
Browse..
< Back
Next >
Cancel

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.



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

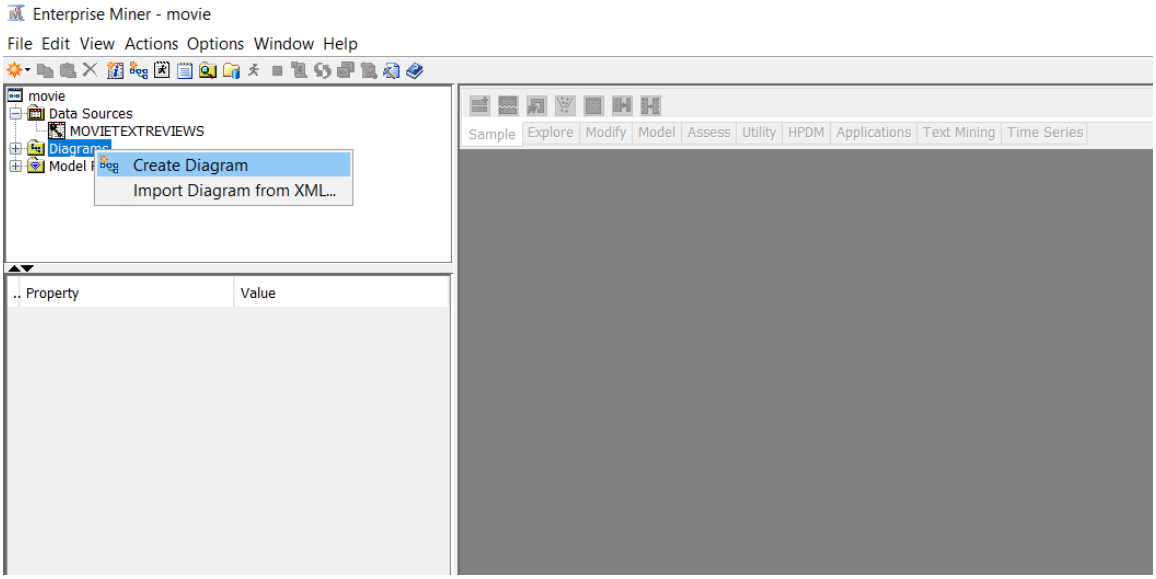
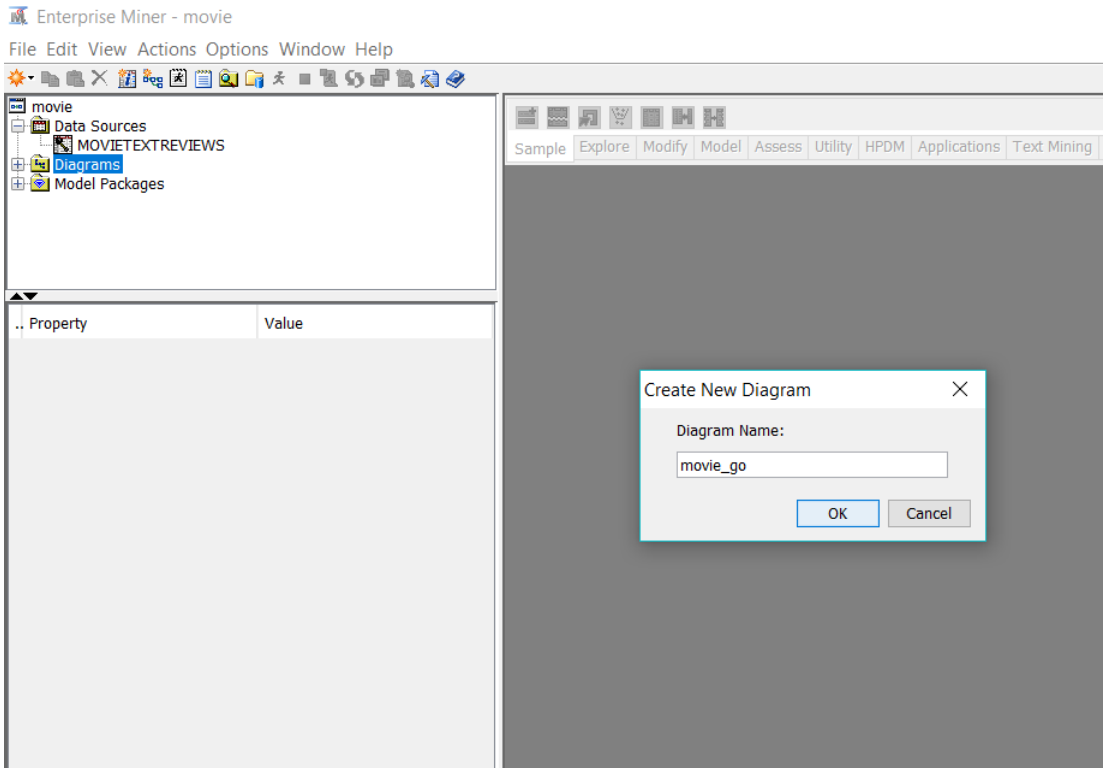
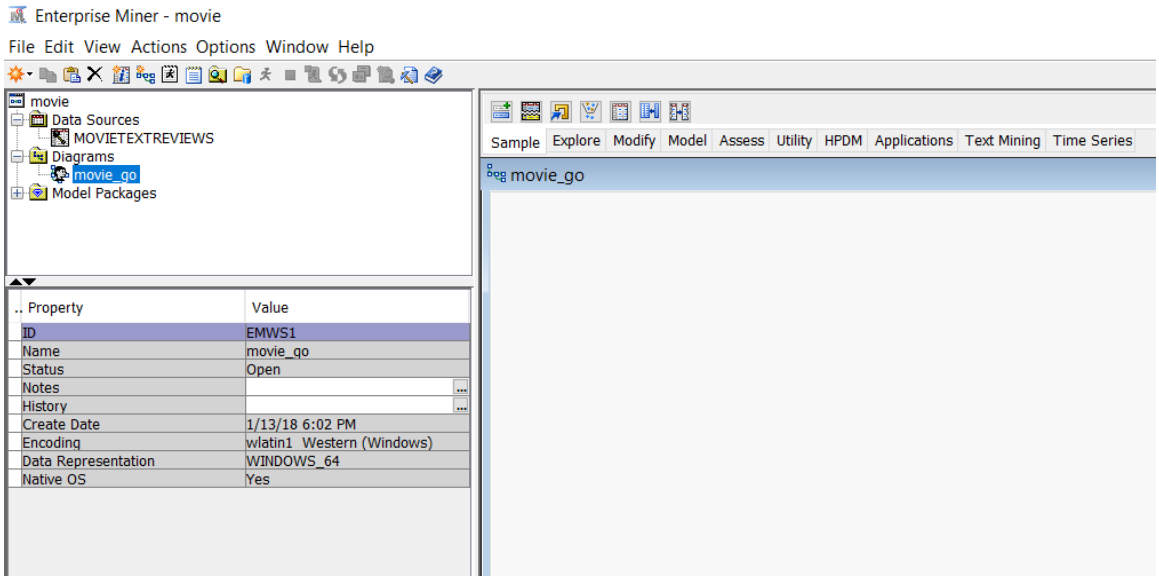


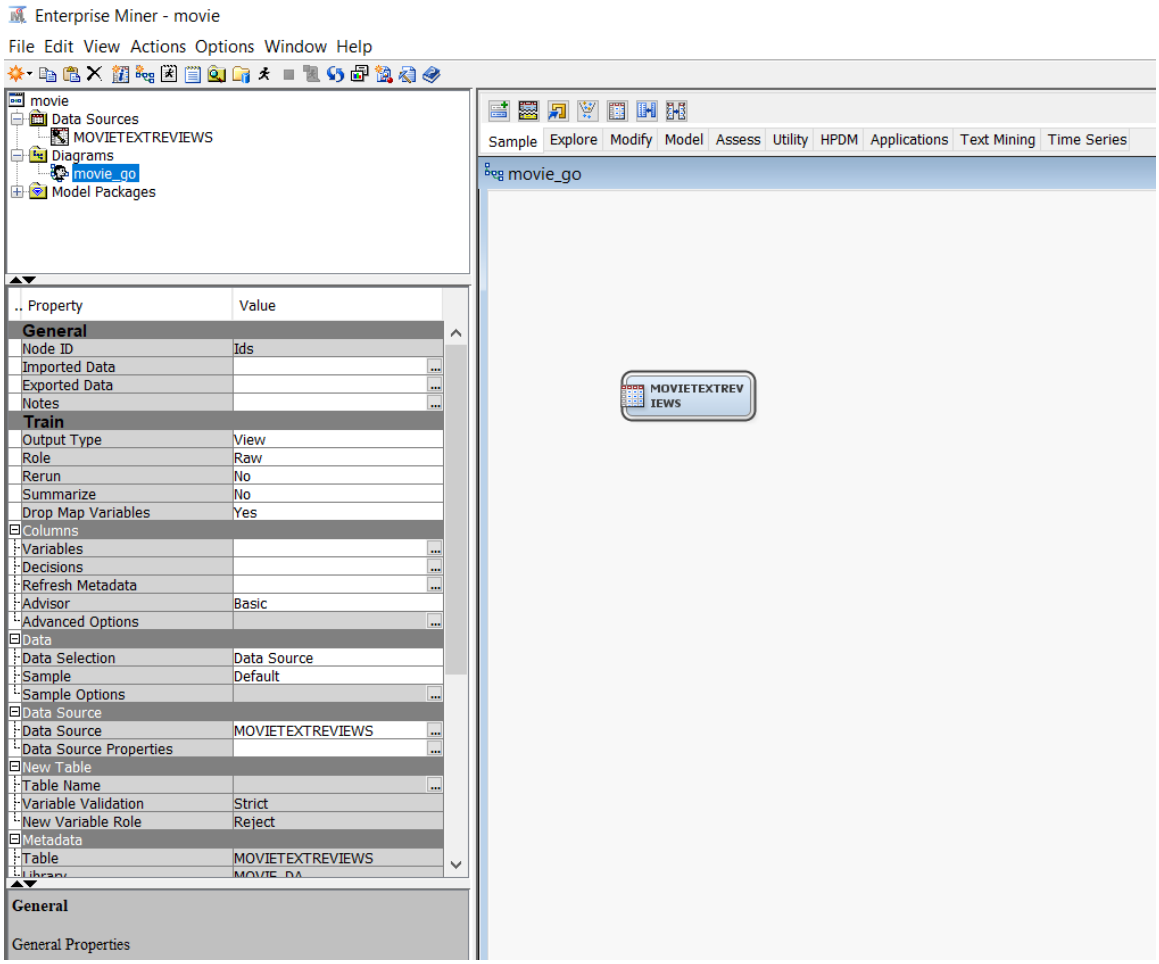
Diagram Name: Enter ‘movie_go.’



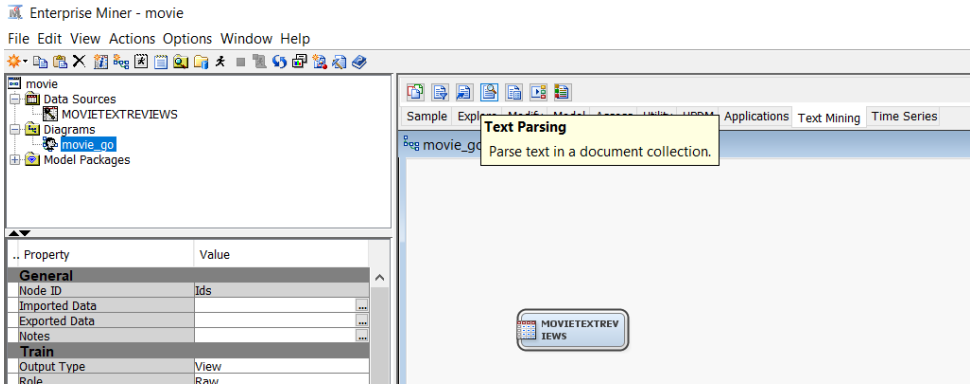
Hit *OK*. Then, you will see a white workspace named “movie_go.”



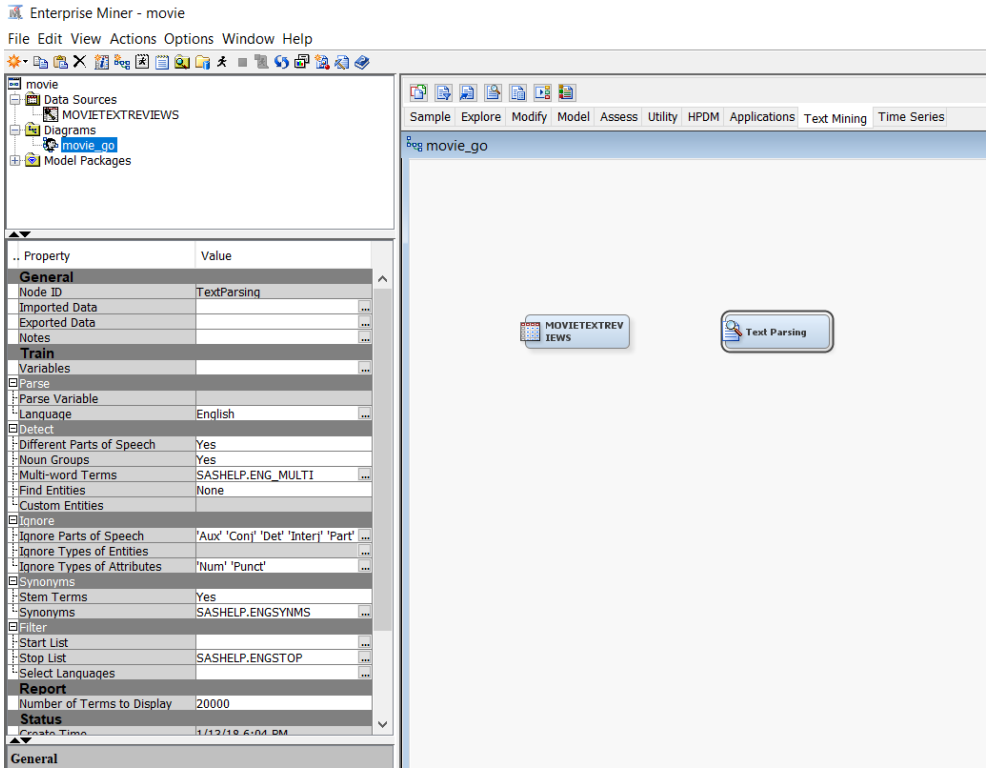
Drag the target data onto the workspace.



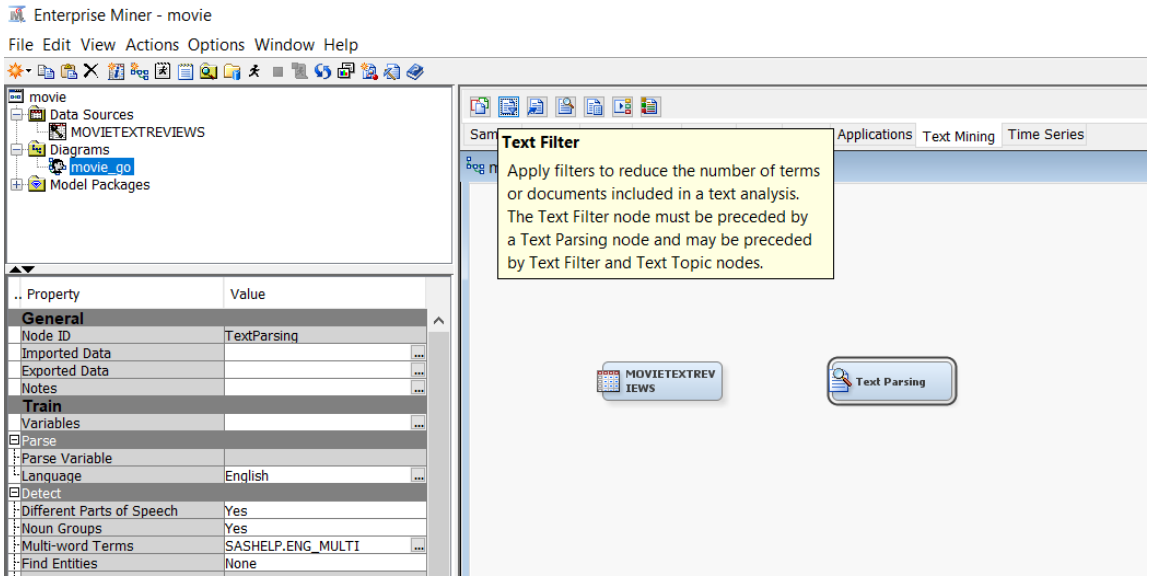
Choose the *Text Mining* tab in the top menu.

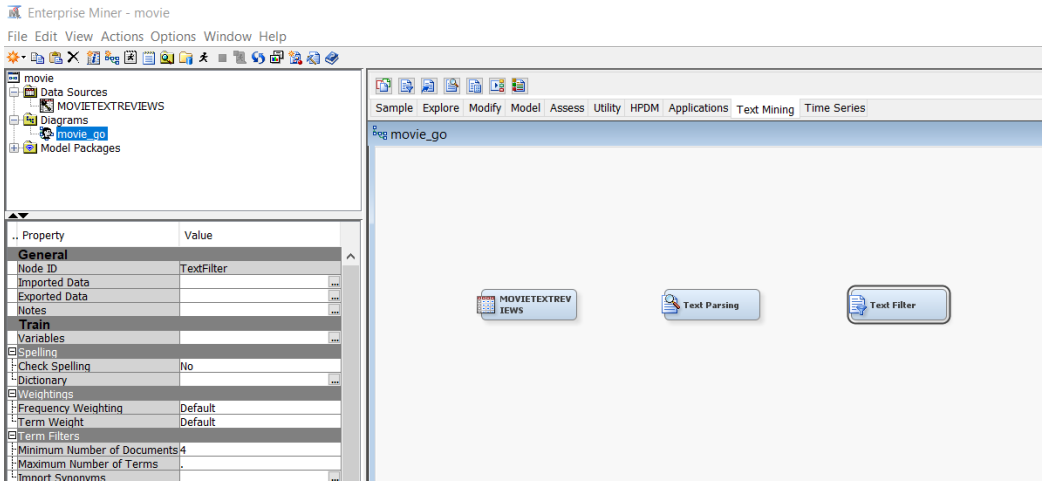


Choose the *Text Parsing* node and drag it onto the workspace.

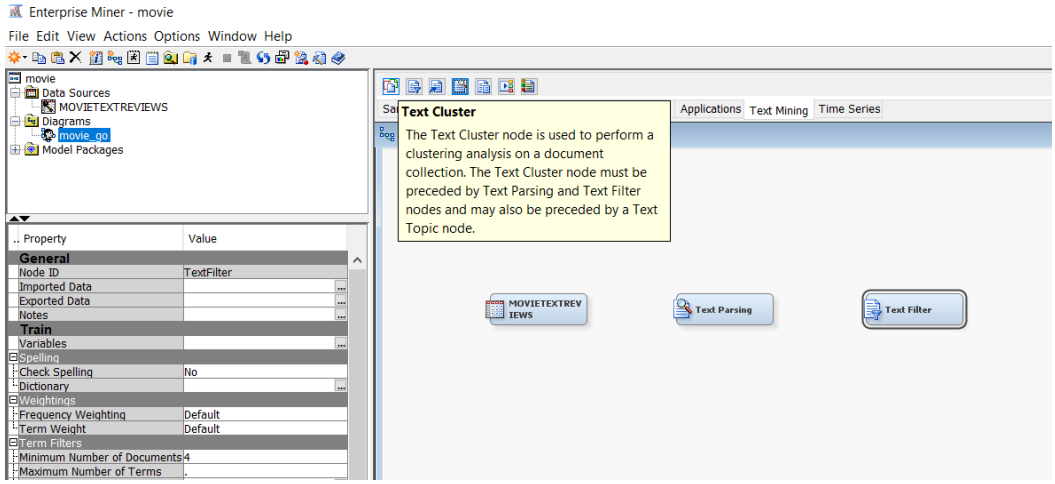


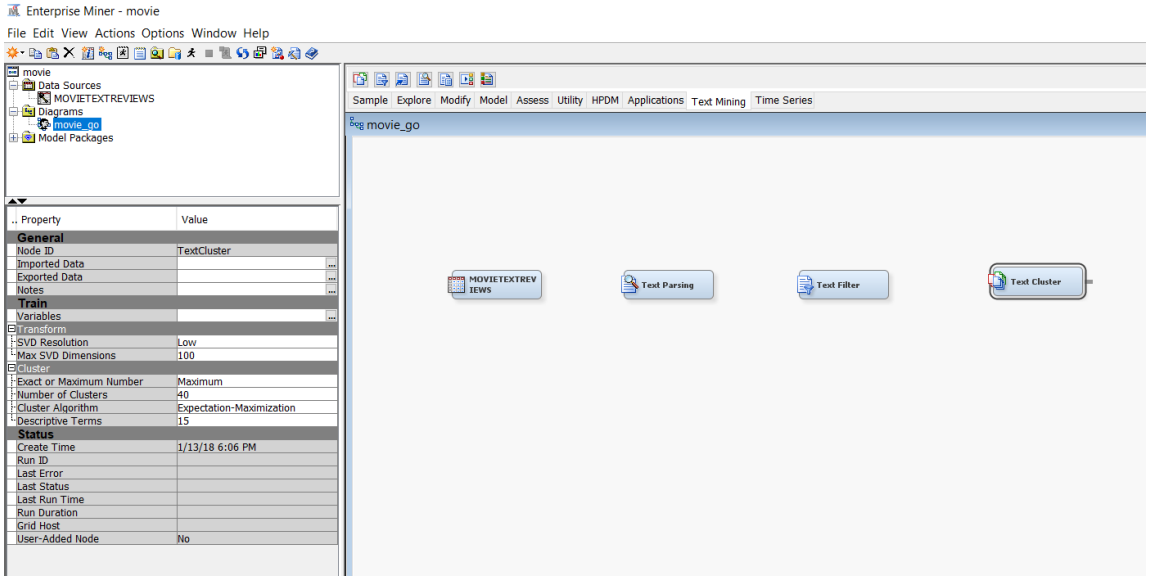
Choose the *Text Filter* node and drag it onto the workspace.



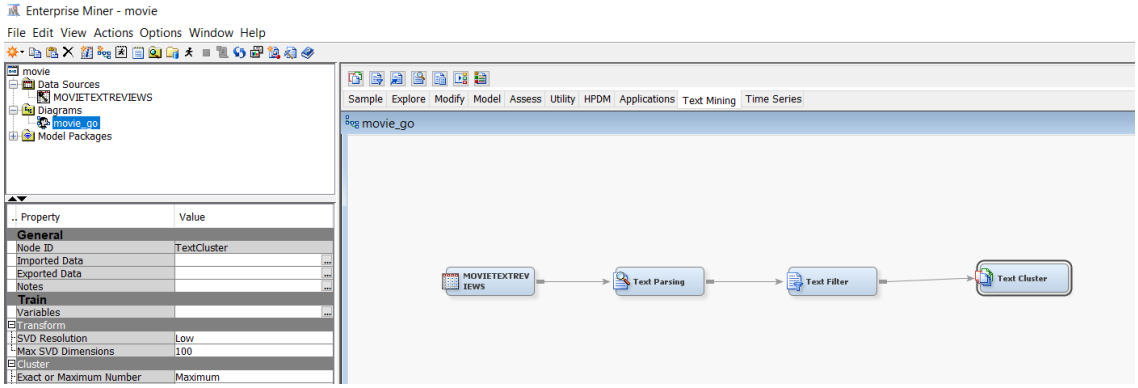


Choose the *Text Cluster* node and drag it onto the workspace.





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,

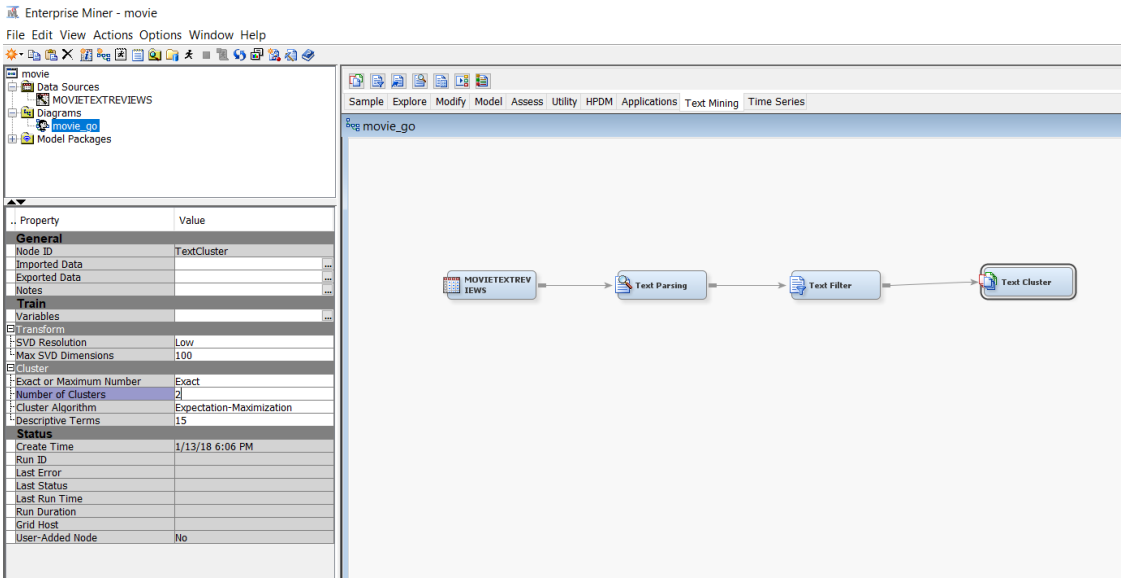
Click the *Text Cluster* node on the workspace. Then, you will see the detailed operation information of the node on the lower left side.

Train < Cluster

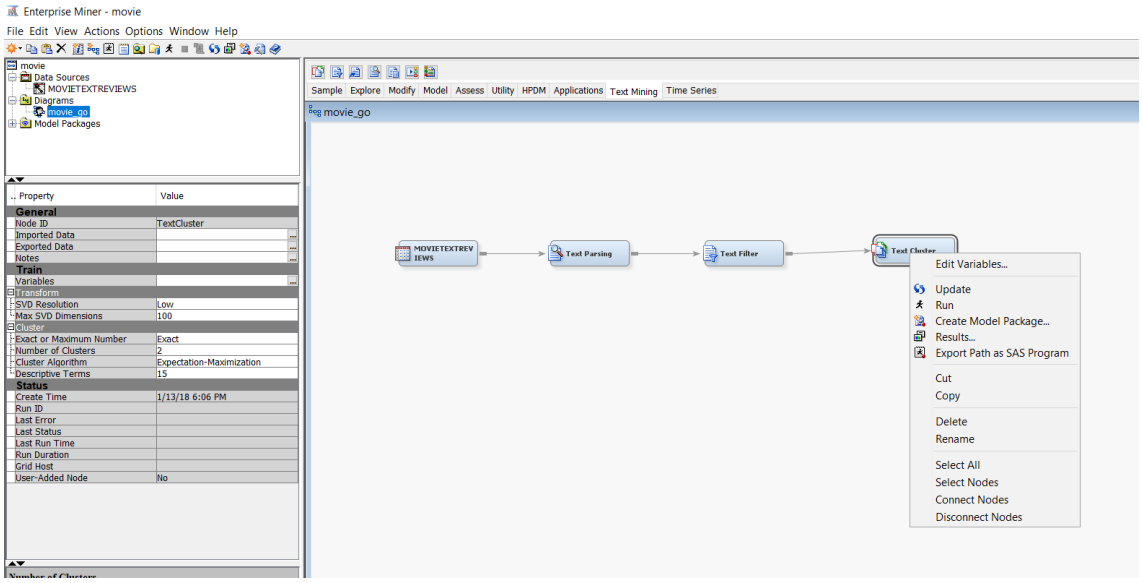
Exact or Maximum Number: Choose '*Exact*.'

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

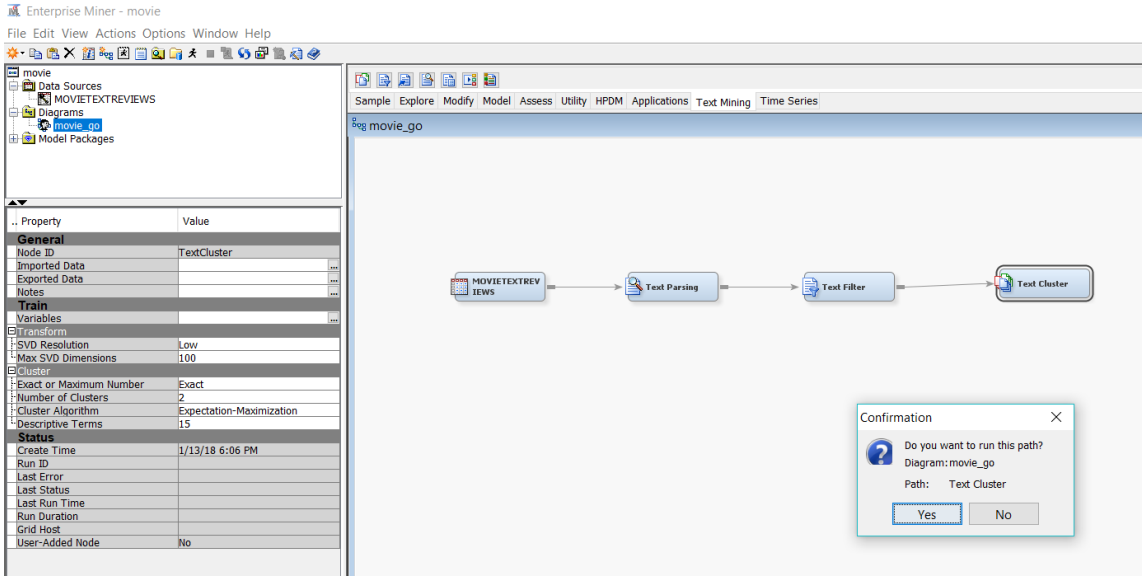
Determine your number of *Descriptive Terms*. (15 is the default option.)



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

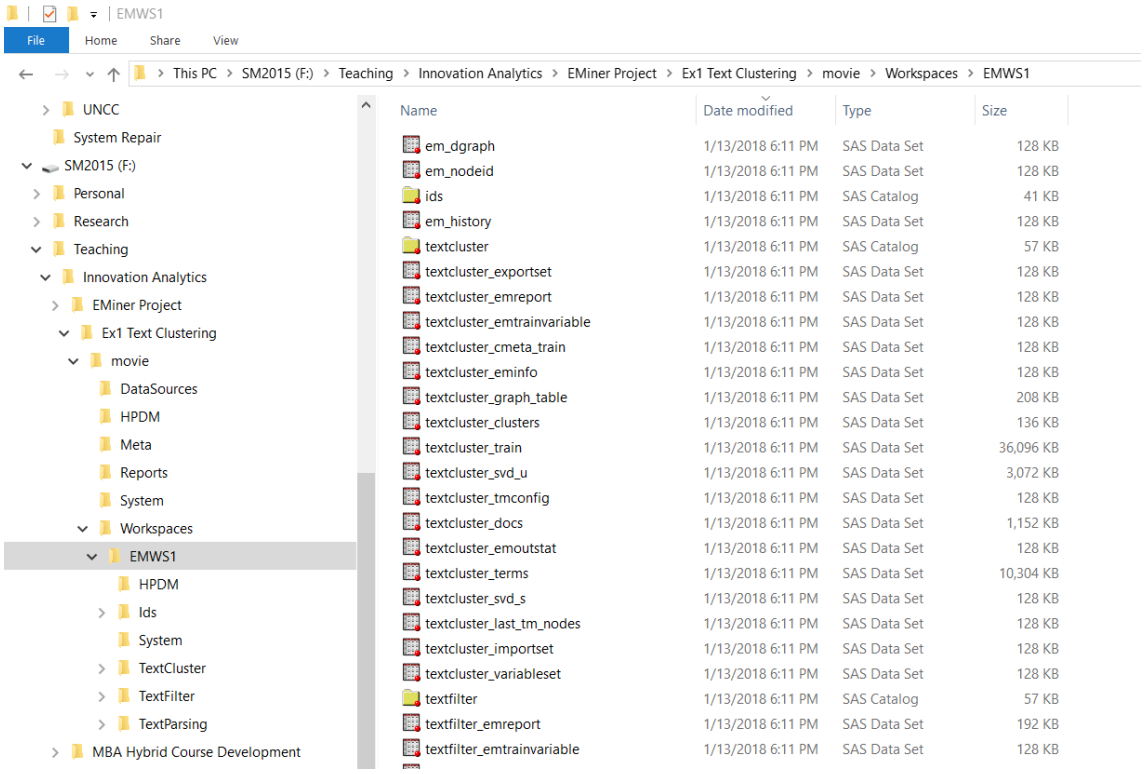


It may take some time to generate the text clustering results (usually, a few minutes).

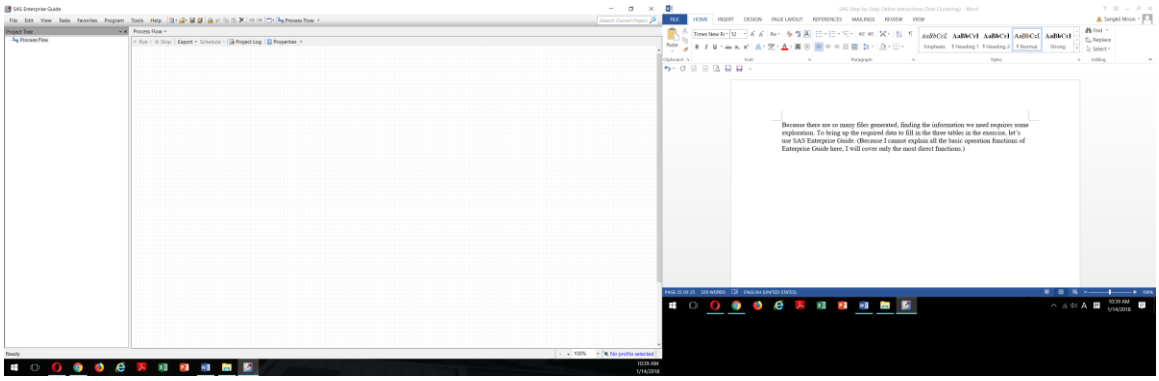


Hit *Yes* to see some results. The results you see here is only a small fraction of the results the program have generated, which can be found in the following directory address:

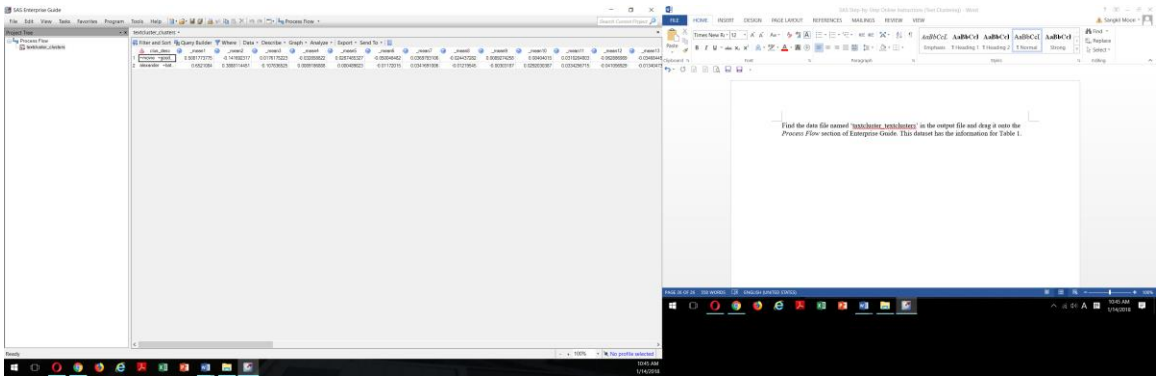
(Your project target directory folder address) < movie < Workspaces < EMWS1



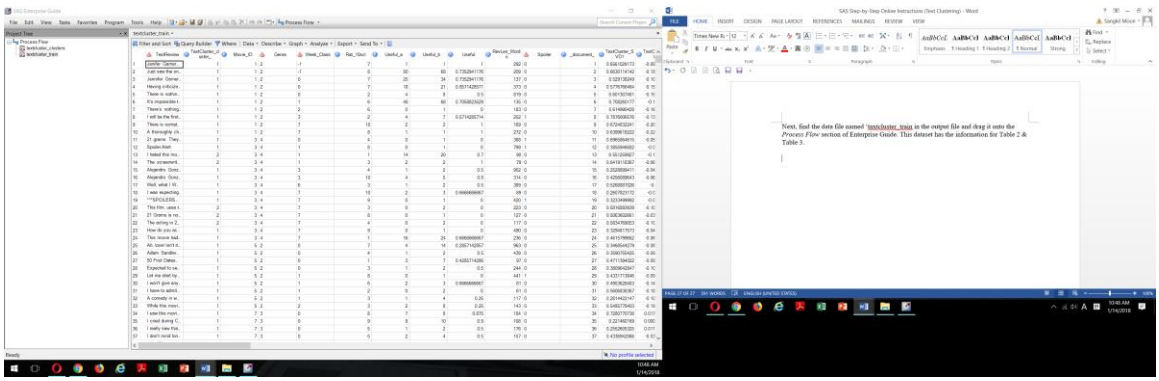
Because there are so many files generated, finding the information we need requires some exploration. To bring up the required data to fill in the three tables in the exercise, let's use *SAS Enterprise Guide*. (Because I cannot explain all the basic operation functions of Enterprise Guide here, I will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



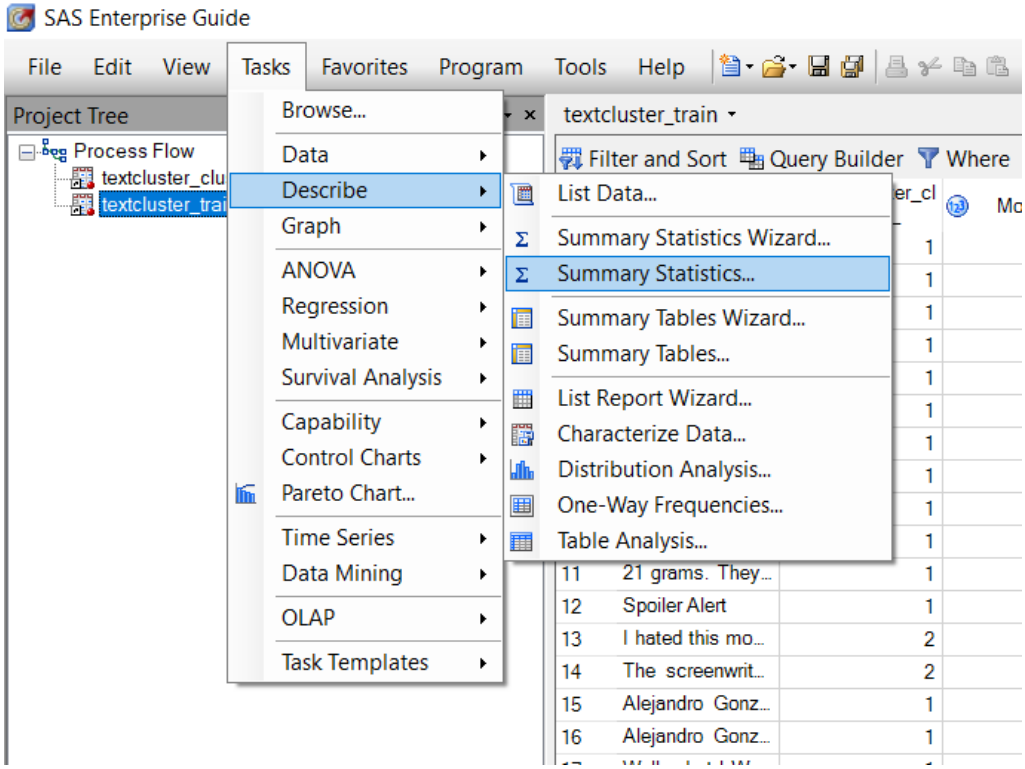
Find the data file named ‘taxtcluster_textclusters’ in the output file and drag it onto the *Process Flow* section of Enterprise Guide. This dataset has the information for Table 1.



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.

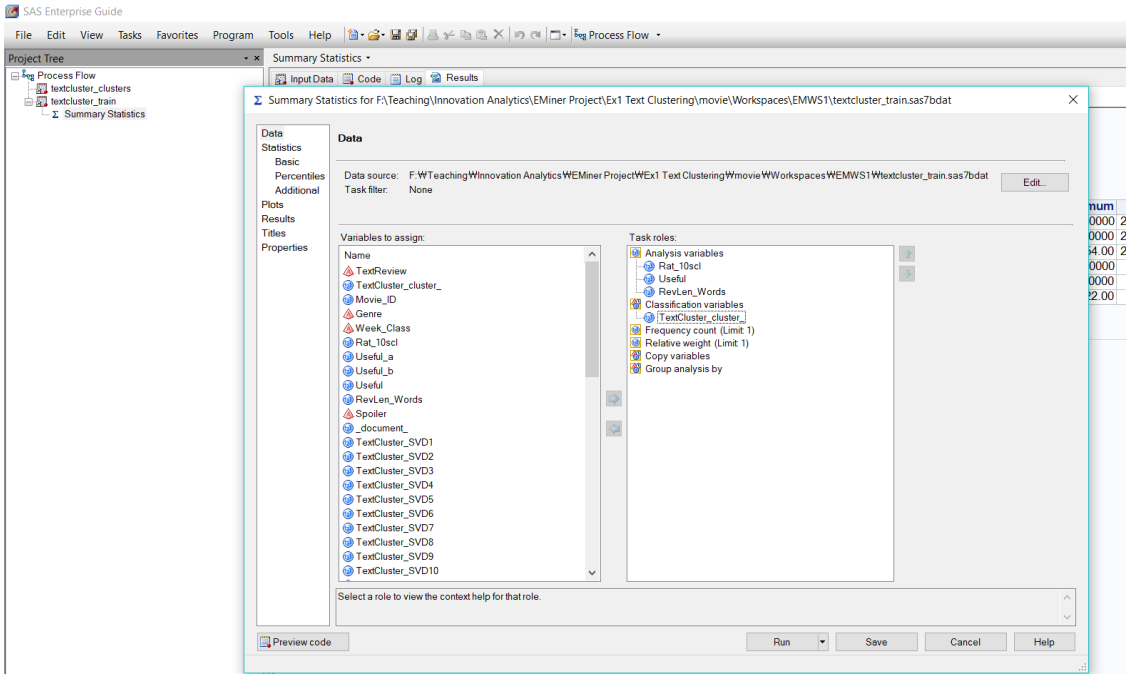


For Table 2 information, choose
Tasks > Describe > Summary Statistics



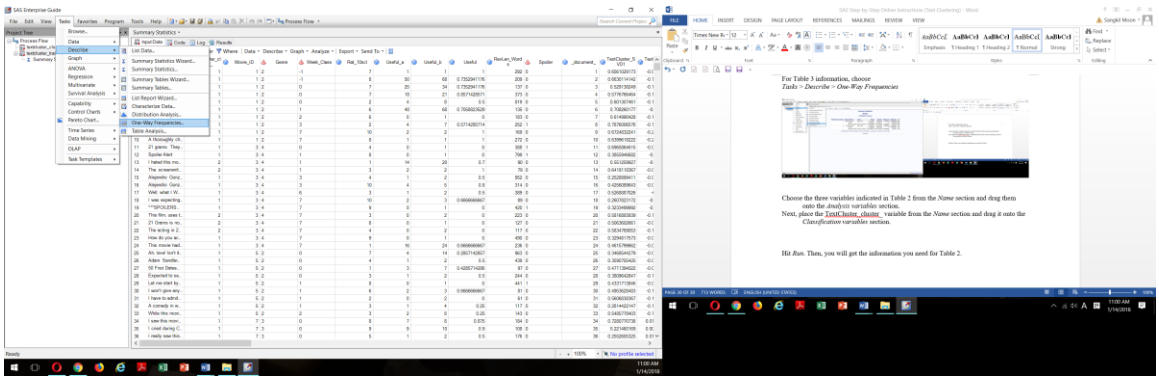
Choose the three variables indicated in Table 2 from the *Name* section and drag them onto the *Analysis variables* section.

Next, place the TextCluster_cluster_ variable from the *Name* section and drag it onto the *Classification variables* section.

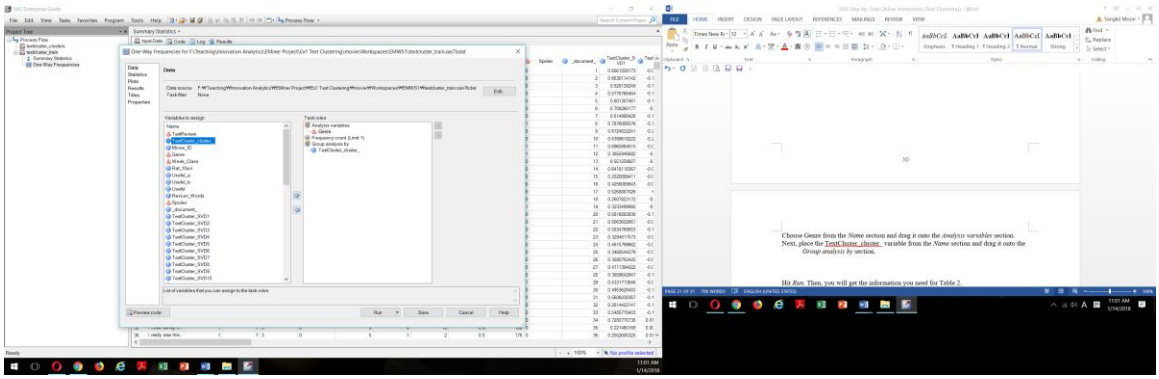


Hit *Run*. Then, you will get the information you need for Table 2.

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



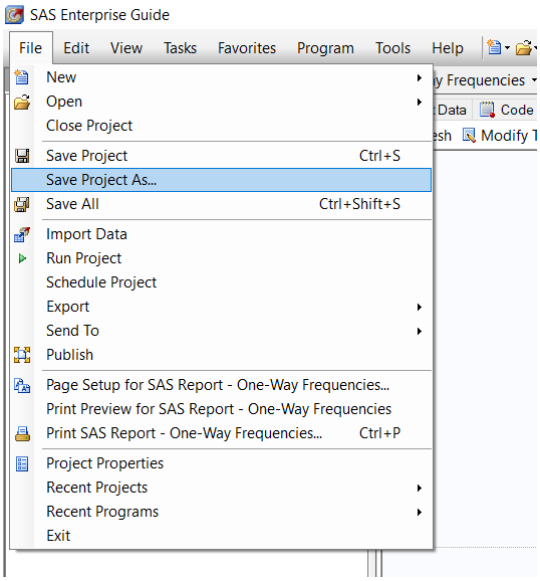
Choose Genre from the *Name* section and drag it onto the *Analysis variables* section. Next, place the TextCluster_cluster_ variable from the *Name* section and drag it onto the *Group analysis by* section.

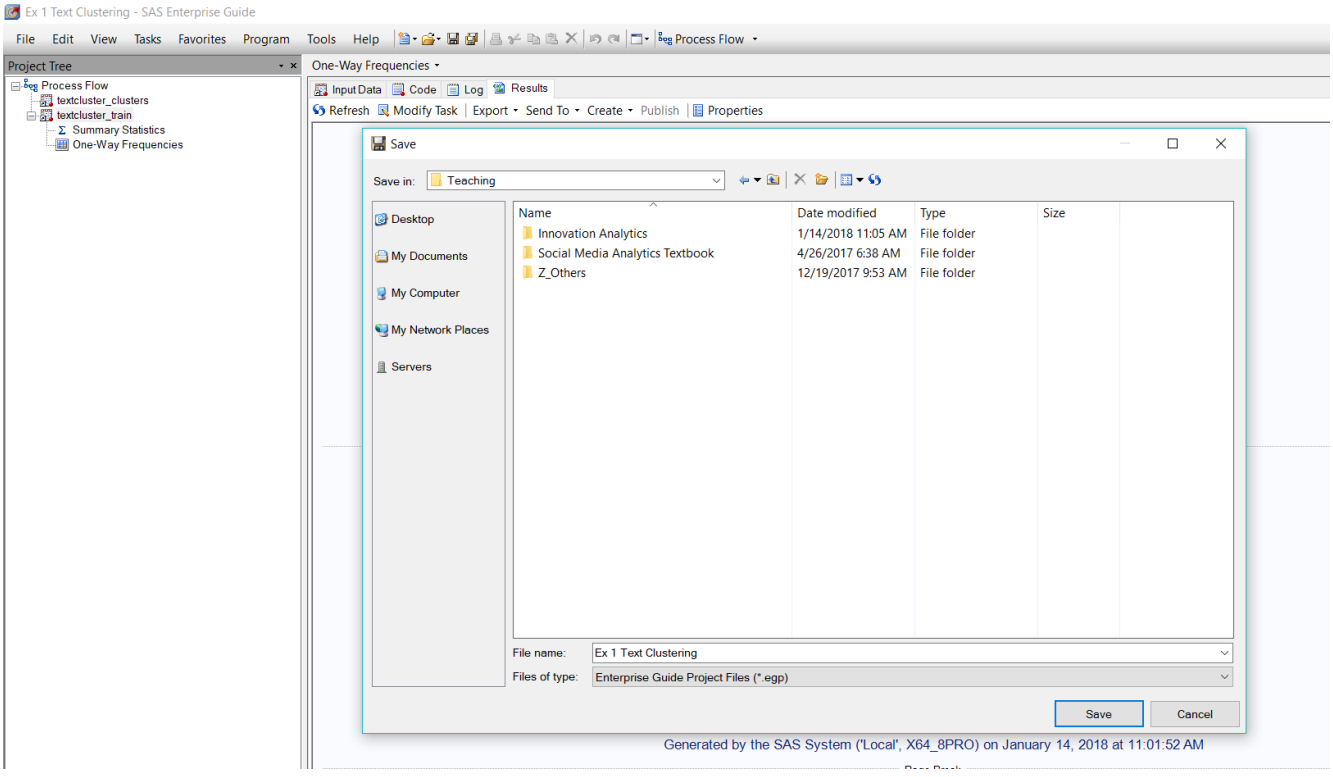


Hit *Run*. Then, you will get the information you need for Table 3.

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

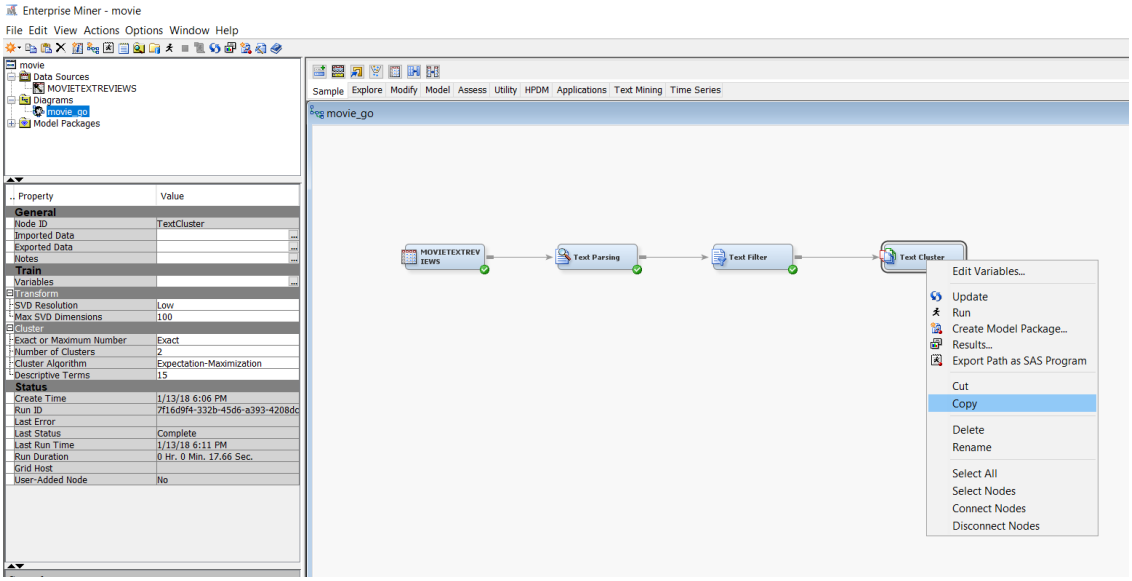
File > Save Project As



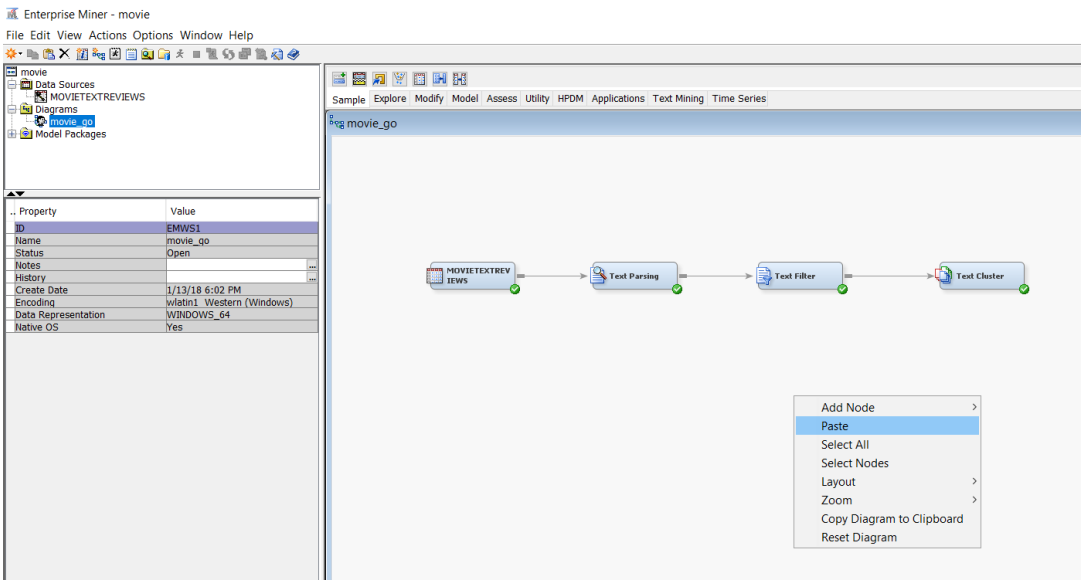


Now, you have obtained all the analysis results for **Task A: Text Clustering – 2 Clusters** in the exercise. You should use necessary information to actually complete Task A.

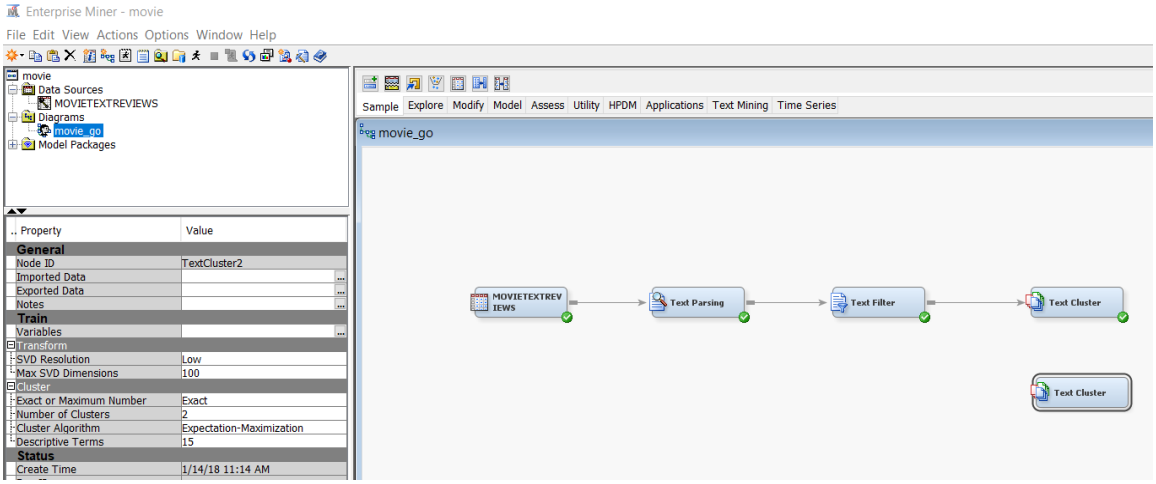
Next, to do **Task B: Text Clustering – 3 Clusters**, go back to the workspace named movie_go on *Enterprise Miner*.
Point to the *Text Cluster* node and choose *Copy* with right click.



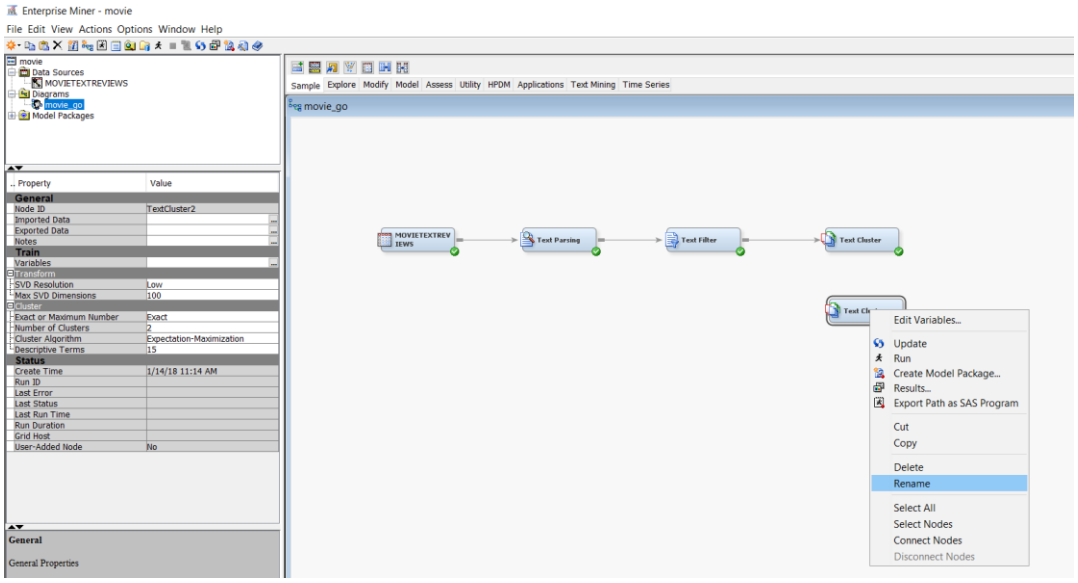
Move the mouse pointer out of the node and place it at any empty space spot and choose *Paste* with right click.



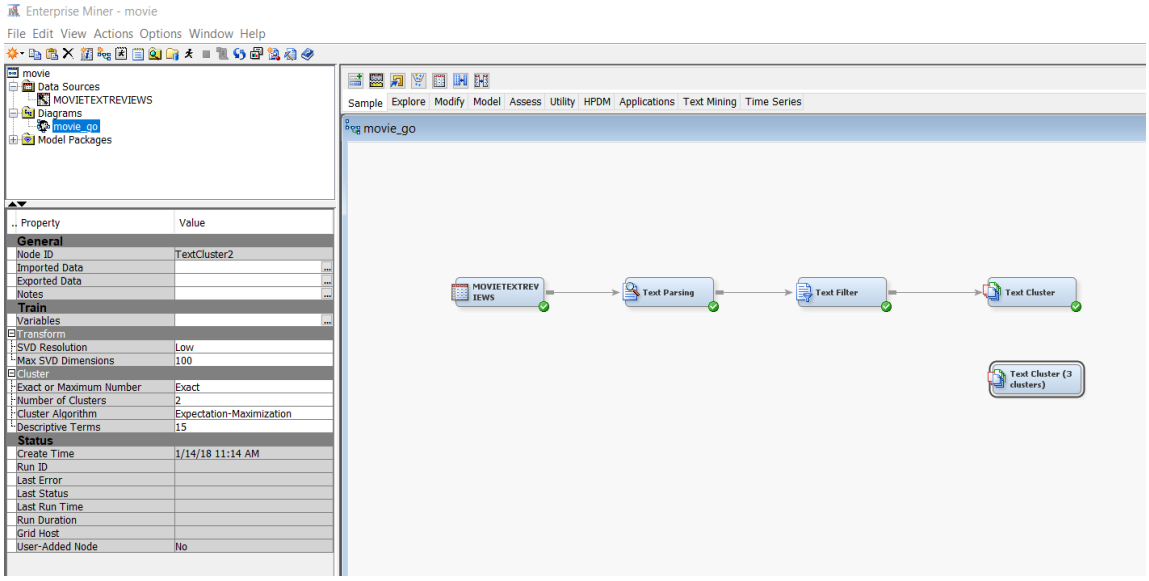
Then, you will get a new *Text Cluster* node.



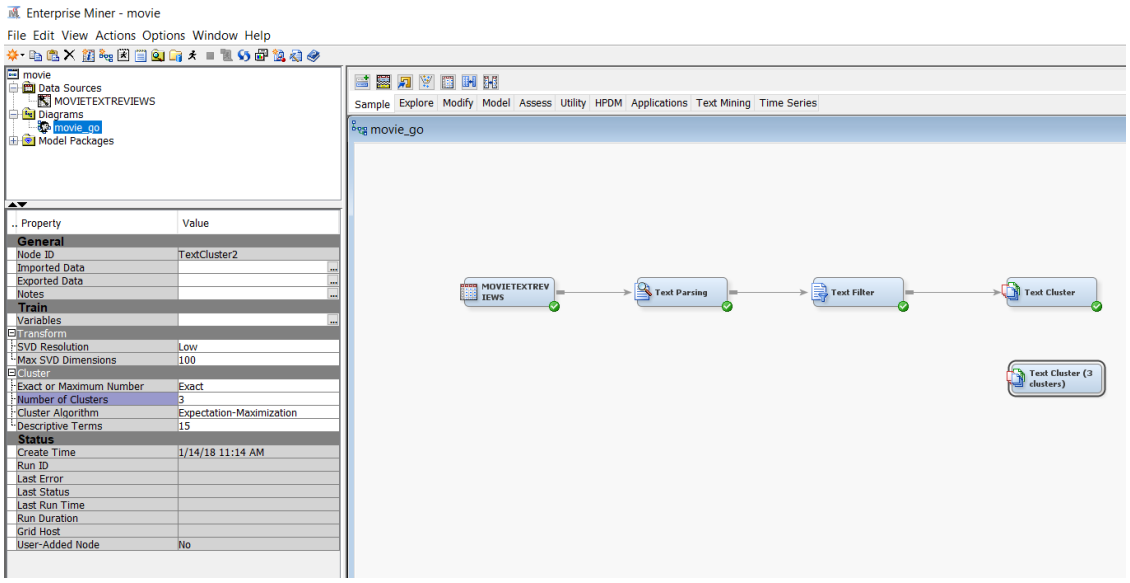
Place the pointer on the new node and choose *Rename* with right click.



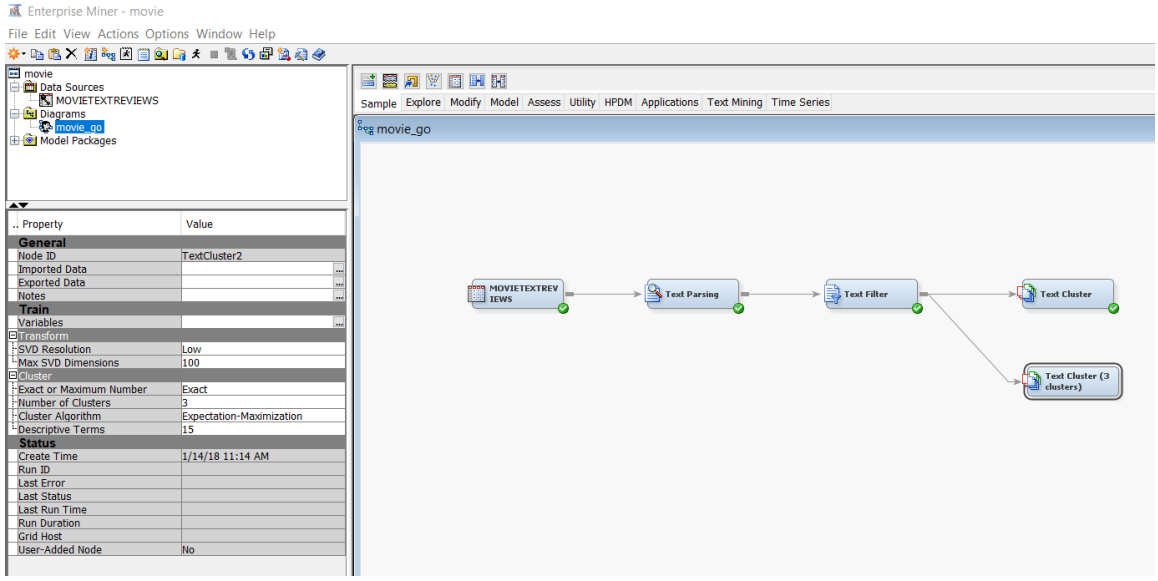
Type in “Text Cluster (3 clusters)” and hit *OK*.



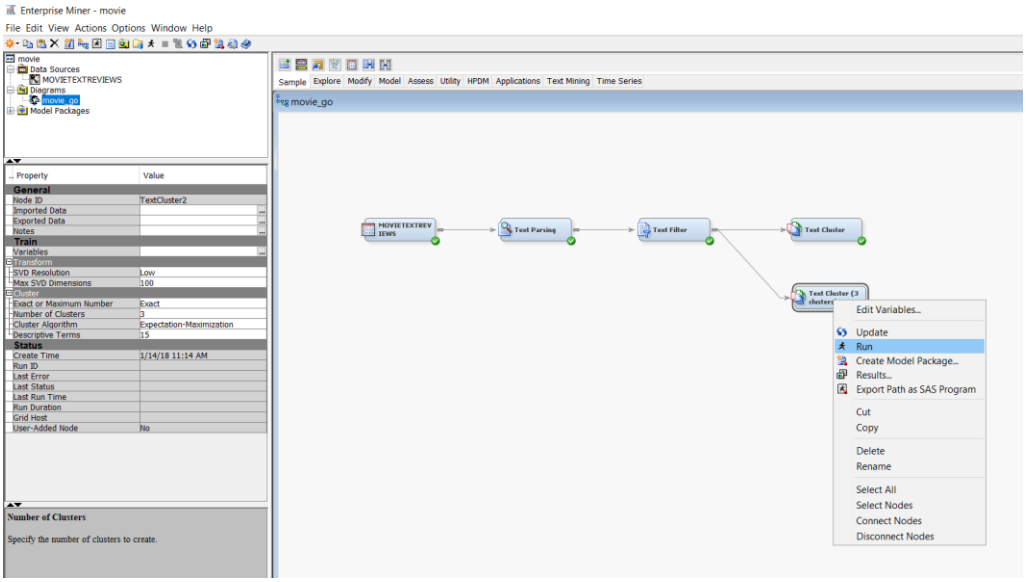
Change *Number of Clusters* from 2 to 3 (at left).



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



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.