

ANLT5070

Unit 3 Assignment 1 Tutorial

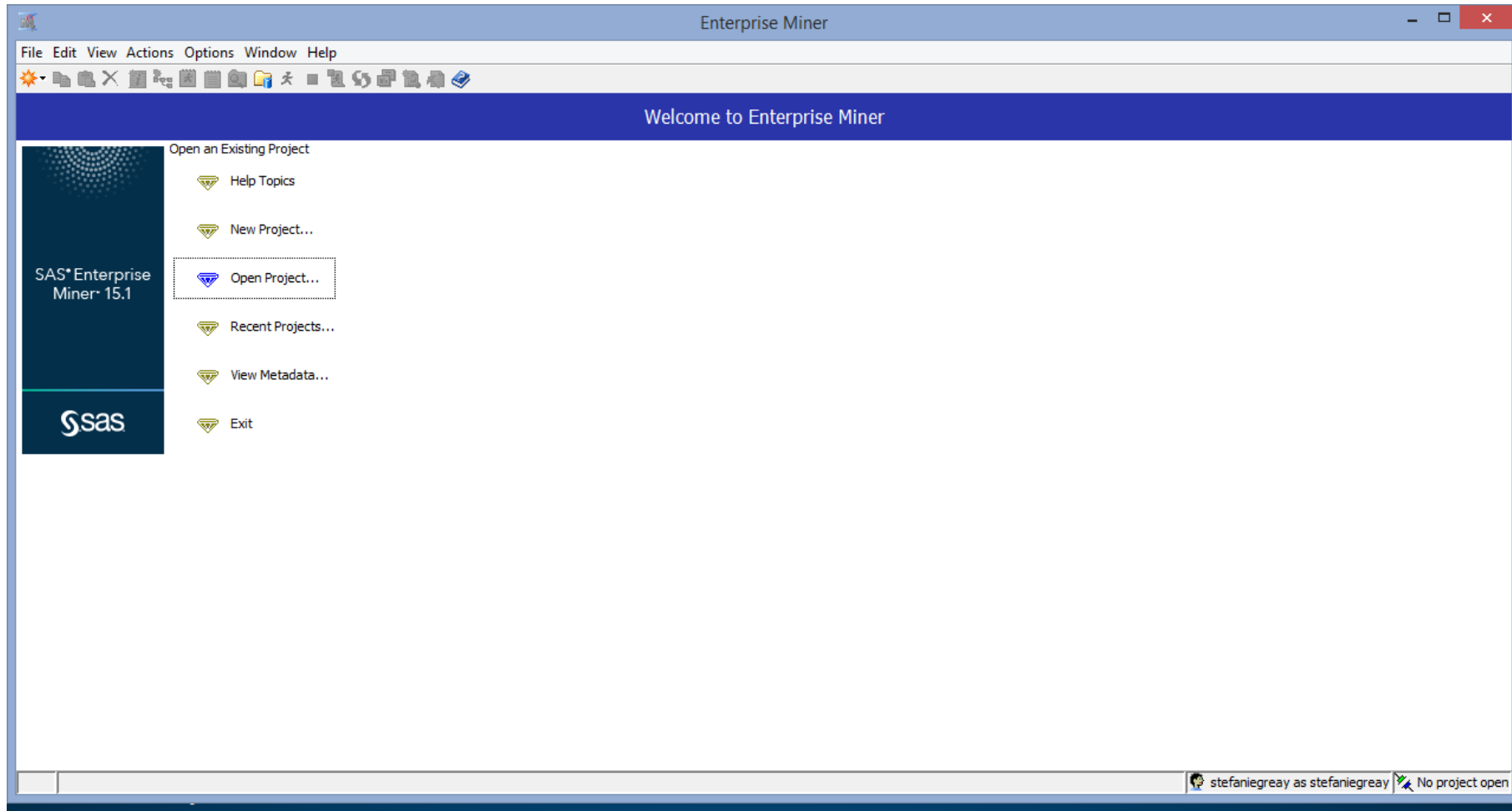


Open project created in Unit 1 and added to in Unit 2

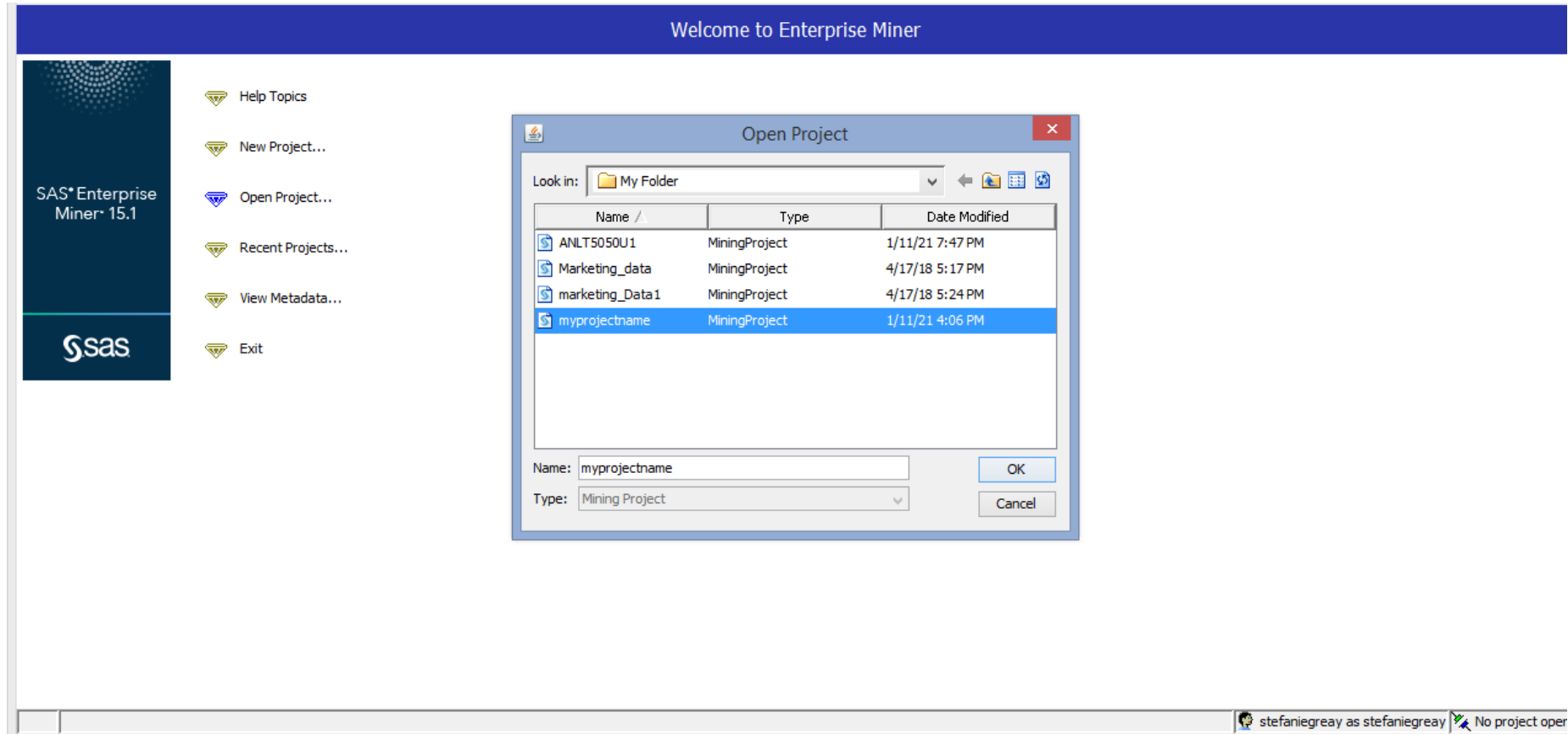
- This assignment is a continuation of the Unit 1 Assignment 1 and Unit 2 Assignment 1 assignments. If you have not successfully completed these, or you have lost your work for them, please re-watch and re-complete those steps first.
- This assignment focuses on the text filtering portion of the text analytics project.



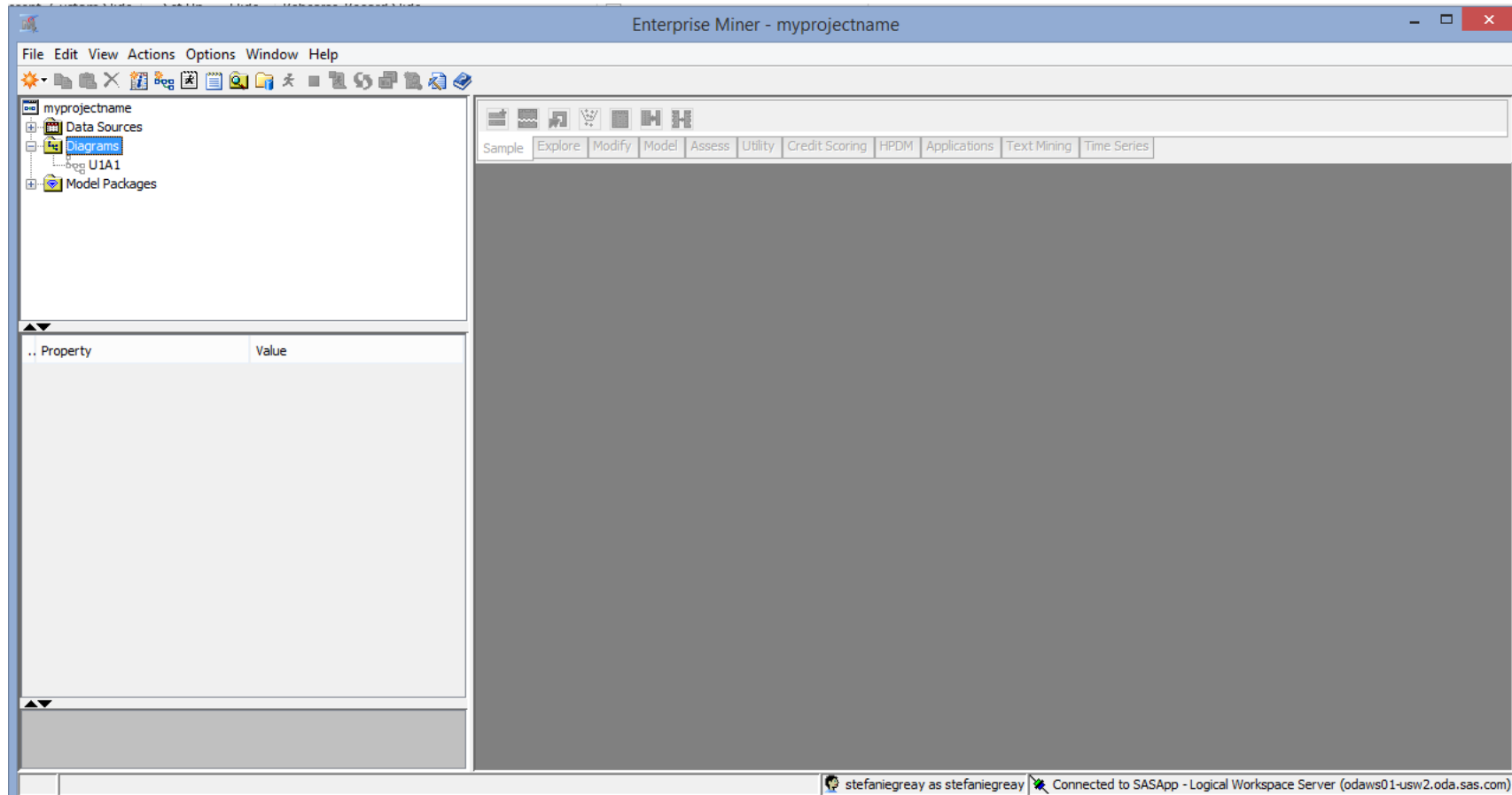
Click on open project



Select the project you created in U1A1, and click “OK” to open the project.



Click on “Diagrams,” navigate to the diagram you created, and double click on it to open it.



Double click on the “Text Filter” node to edit the options for filtering the text data.

The screenshot displays the SAS Enterprise Miner interface. On the left, a project tree shows 'myprojectname' with sub-items 'Data Sources', 'Diagrams', and 'Model Packages'. The 'Diagrams' folder is expanded, showing 'U1A1'. Below the tree is a property window for the 'TextFilter' node, showing various configuration options under 'General', 'Train', 'Spelling', 'Weightings', 'Term Filters', and 'Dictionary'. The main workspace shows a workflow diagram with four nodes: 'Text Import', 'Text Parsing', 'Text Filter', and 'SAS Code'. Arrows indicate the flow from 'Text Import' to 'Text Parsing' to 'Text Filter', and from 'SAS Code' to 'Text Filter'. The status bar at the bottom indicates 'Diagram U1A1 opened' and 'Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)'.

Property	Value
General	
Node ID	TextFilter
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Spelling	
Check Spelling	Yes
Dictionary	
Weightings	
Frequency Weighting	Default
Term Weight	Default
Term Filters	
Minimum Number of Documents	4
Maximum Number of Terms	.
Import Synonyms	
Dictionary	
Specify a data set of correctly spelled terms	



Leave the Frequency Weighting and Term Weight as default, and change the Minimum Number of Documents to 2 and Maximum Number of Terms to 100.

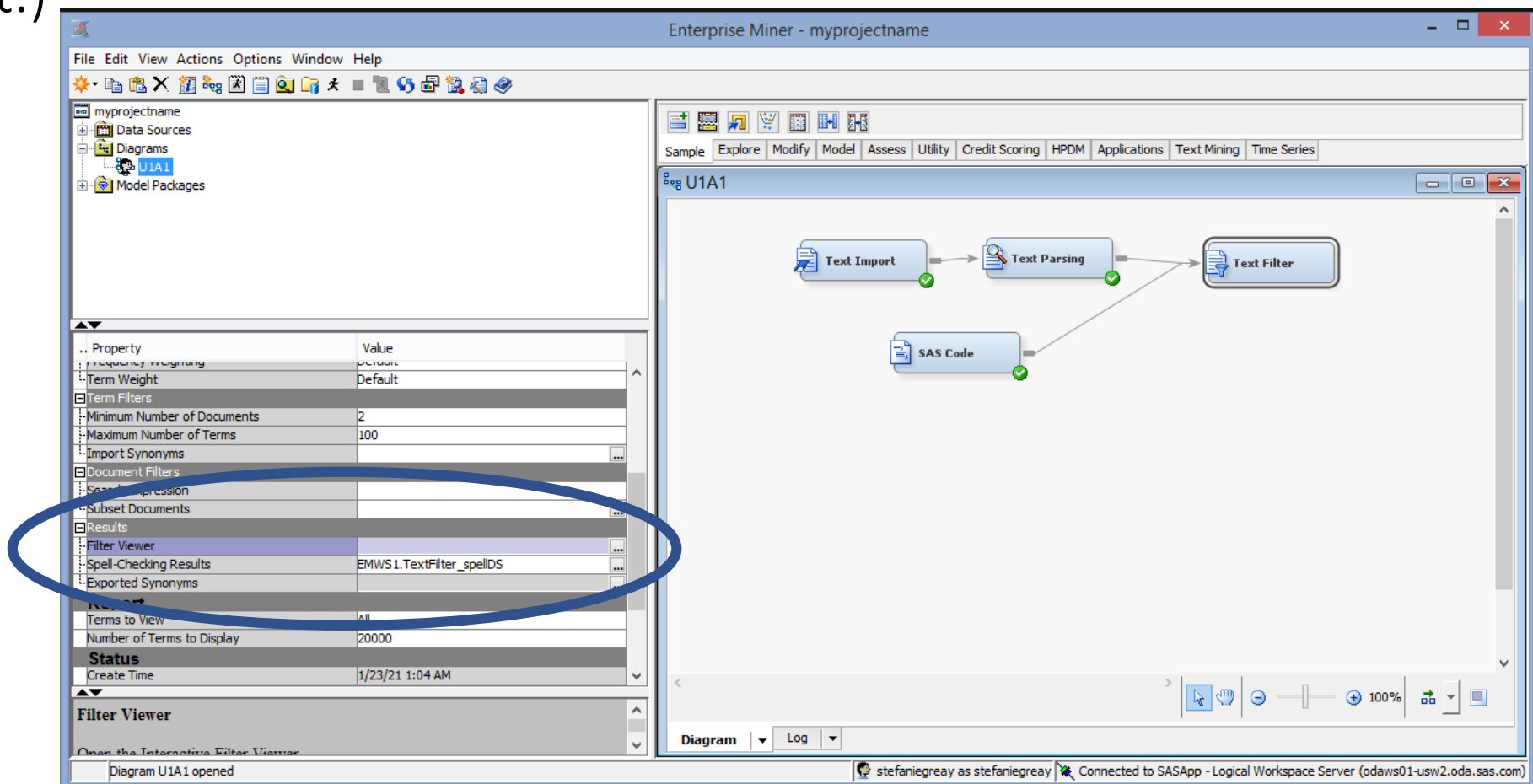
The screenshot displays the SAS Enterprise Miner interface. On the left, a tree view shows the project structure with 'myprojectname' expanded, containing 'Data Sources', 'Diagrams', 'U1A1', and 'Model Packages'. Below this, a properties table for the 'TextFilter' node is shown, with a blue oval highlighting the 'Term Filters' section.

Property	Value
General	
Node ID	TextFilter
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Spelling	
Check Spelling	Yes
Dictionary	MYDATA.ENGDICT
Weightings	
Frequency Weighting	Default
Term Weight	Default
Term Filters	
Minimum Number of Documents	2
Maximum Number of Terms	100
Import Synonyms	

On the right, the 'U1A1' diagram window shows a workflow with four nodes: 'Text Import', 'Text Parsing', 'Text Filter', and 'SAS Code'. Arrows indicate the flow from 'Text Import' to 'Text Parsing' to 'Text Filter', and from 'SAS Code' to 'Text Filter'. All nodes have a green checkmark, indicating they are active or successful.



You can review the filters interactively by clicking on the ellipses (3 dots) next to “Filter Viewer.” (Note that you must run the Text Filter node first.)



Check or uncheck the “keep” checkbox for any filters you think should be added or removed.

Interactive Filter Viewer

File Edit View Window

Search : Apply Clear

Documents

TEXT	ACCESSED	CREATED	EXTENSI...	FILTERED	FILTE...	LANGUAGE	MODIFIED	NAME	OMITTED
1 Paper 10360-2016 Nine Frequently Asked Questions about Getting Started with SAS® Visual	2021-01-1...	2021-01-1...	.pdf	/home/ste...	54492.0	English	2021-01-1...	10360-201...	0.0
1 Paper 10401-2016 Responsible Gambling Model at Veikkaus Tero Kallioniemi, Veikkaus Oy	2021-01-1...	2021-01-1...	.pdf	/home/ste...	15834.0	English	2021-01-1...	10401-201...	0.0
1 Paper 10600-2016 You Can Bet On It, The Missing Rows are Preserved with PRELOADFMT	2021-01-1...	2021-01-1...	.pdf	/home/ste...	16612.0	English	2021-01-1...	10600-201...	0.0
1 Paper 10740-2016 Developing an On-Demand Web Report Platform Using Stored Processes	2021-01-1...	2021-01-1...	.pdf	/home/ste...	22946.0	English	2021-01-1...	10740-201...	0.0
1 Paper 11221-2016 Leads and Lags: Static and Dynamic Queues in the SAS® DATA STEP	2021-01-1...	2021-01-1...	.pdf	/home/ste...	29382.0	English	2021-01-1...	11221-201...	0.0
1 Paper 11775 - 2016 Using SAS® Text Miner for Automatic Categorization of Blog Posts on a	2021-01-1...	2021-01-1...	.pdf	/home/ste...	18223.0	English	2021-01-1...	11775-201...	0.0
1 Paper 2101-2016 Using the Kaplan-Meier Product-Limit Estimator to Adjust NFL Yardage	2021-01-1...	2021-01-1...	.pdf	/home/ste...	46710.0	English	2021-01-1...	2101-2016...	0.0
1 Paper 2480-2016 Performing Pattern Matching by Using Perl Regular Expressions Arthur Li,	2021-01-1...	2021-01-1...	.pdf	/home/ste...	33364.0	English	2021-01-1...	2480-2016...	0.0
3940: Fantasizing about the Big Data of NFL Fantasy Football, or Time to Get a Life Clint	2021-01-1...	2021-01-1...	.pdf	/home/ste...	199.0	English	2021-01-1...	3940-2016...	0.0
SAS 5580-2016 MACRO VARIABLES IN SAS® ENTERPRISE GUIDE® Khoi To, Office of Planning	2021-01-1...	2021-01-1...	.pdf	/home/ste...	9118.0	English	2021-01-1...	5580-2016...	0.0
1 SAS 5581-2016 USING PROC TABULATE AND LAG(n) FUNCTION FOR RATES OF CHANGE	2021-01-1...	2021-01-1...	.pdf	/home/ste...	10957.0	English	2021-01-1...	5581-2016...	0.0
6500: Research Problems Arising in Sports Statistics Tim Swartz Content for this session has	2021-01-1...	2021-01-1...	.pdf	/home/ste...	163.0	English	2021-01-1...	6500-2016...	0.0
1 Paper: 7020-2016 Three Methods to Dynamically Assign Colors to Plots Based on Group	2021-01-1...	2021-01-1...	.pdf	/home/ste...	4682.0	English	2021-01-1...	7020-2016...	0.0

Terms

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
product	194	28	<input checked="" type="checkbox"/>	0.33		Alpha
program	169	21	<input checked="" type="checkbox"/>	0.391		Alpha
option	240	20	<input checked="" type="checkbox"/>	0.568		Alpha
output	205	20	<input checked="" type="checkbox"/>	0.372		Alpha
add	194	19	<input checked="" type="checkbox"/>	0.403		Alpha
generate	113	17	<input checked="" type="checkbox"/>	0.415		Alpha
version	61	17	<input checked="" type="checkbox"/>	0.473		Alpha
file	261	16	<input checked="" type="checkbox"/>	0.439		Alpha
email	75	15	<input checked="" type="checkbox"/>	0.442		Alpha
global	167	14	<input checked="" type="checkbox"/>	0.667		Alpha
line	87	14	<input checked="" type="checkbox"/>	0.453		Alpha
tab	67	12	<input checked="" type="checkbox"/>	0.507		Alpha



Navigate to the Term-by-Document Matrix by clicking on the ellipses (3 dots) next to “Exported Data” in the Text Filter Node.

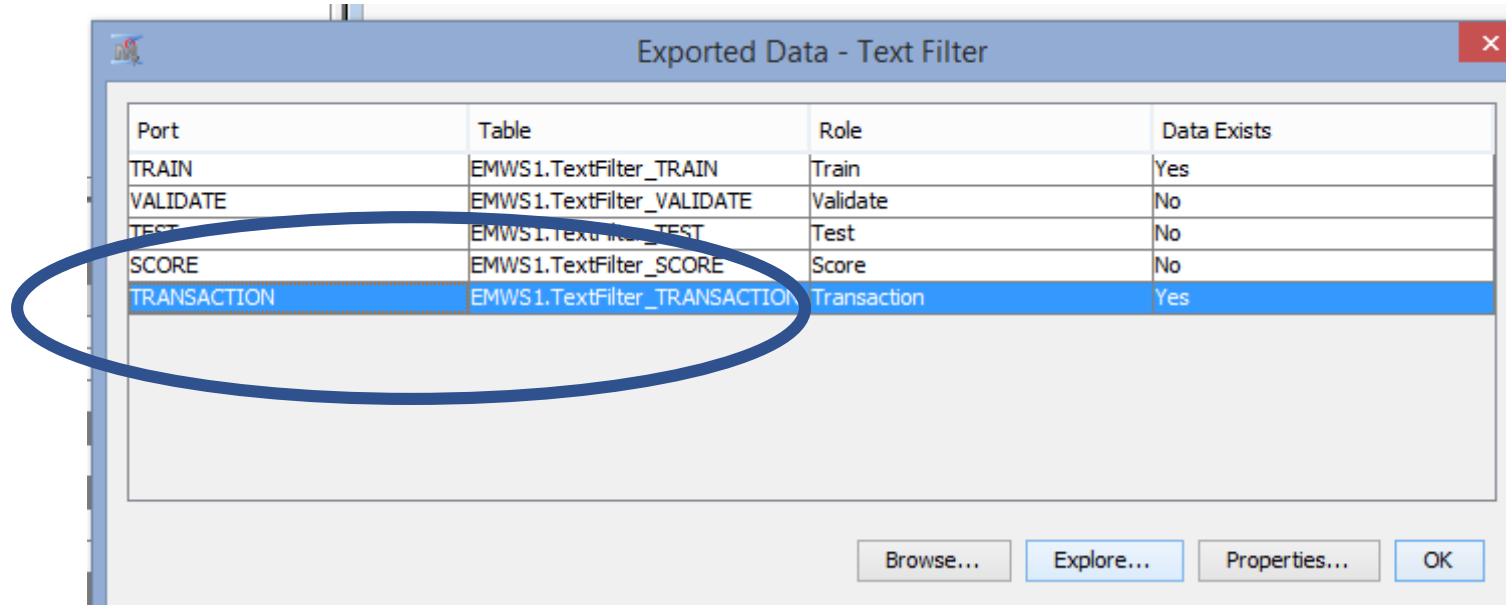
The screenshot displays the Enterprise Miner application window titled "Enterprise Miner - myprojectname". The interface is divided into several panes:

- Left Pane:** A tree view showing the project structure with "myprojectname", "Data Sources", "Diagrams", "U1A1", and "Model Packages".
- Bottom-Left Pane:** A properties table for the selected node. The "General" tab is active, showing the "Node ID" as "TextFilter". The "Exported Data" section is highlighted with a blue oval, indicating the area to click on to access the Term-by-Document Matrix.
- Right Pane:** A workflow diagram titled "U1A1" showing a sequence of steps: "Text Import" → "Text Parsing" → "Text Filter". A "SAS Code" node is also connected to the "Text Filter" node. All steps have green checkmarks indicating they are completed or successful.

The "Exported Data" section in the properties table is highlighted with a blue oval, indicating the area to click on to access the Term-by-Document Matrix.



Click on the “TRANSACTION” dataset and click “Explore.”



The Term-by-Document dataset/matrix is shown in the bottom pane. You can also find and download this dataset using code or in SAS Studio.

The screenshot shows the SAS Studio interface for the dataset `EMWS1.TextFilter_TRANSACTION`. The interface is divided into three main panes:

- Sample Properties:** A table showing dataset metadata.
- Sample Statistics:** A table showing summary statistics for the first four variables.
- EMWS1.TextFilter_TRANSACTION:** A data table showing the first 15 observations. This pane is circled in blue.

Sample Properties

Property	Value
Rows	Unknown
Columns	4
Library	EMWS1
Member	TEXTFILTER_TRANSACTION
Type	VIEW
Sample Method	Top
Fetch Size	Default
Fetch Rows	922
Random Seed	12345

Sample Statistics

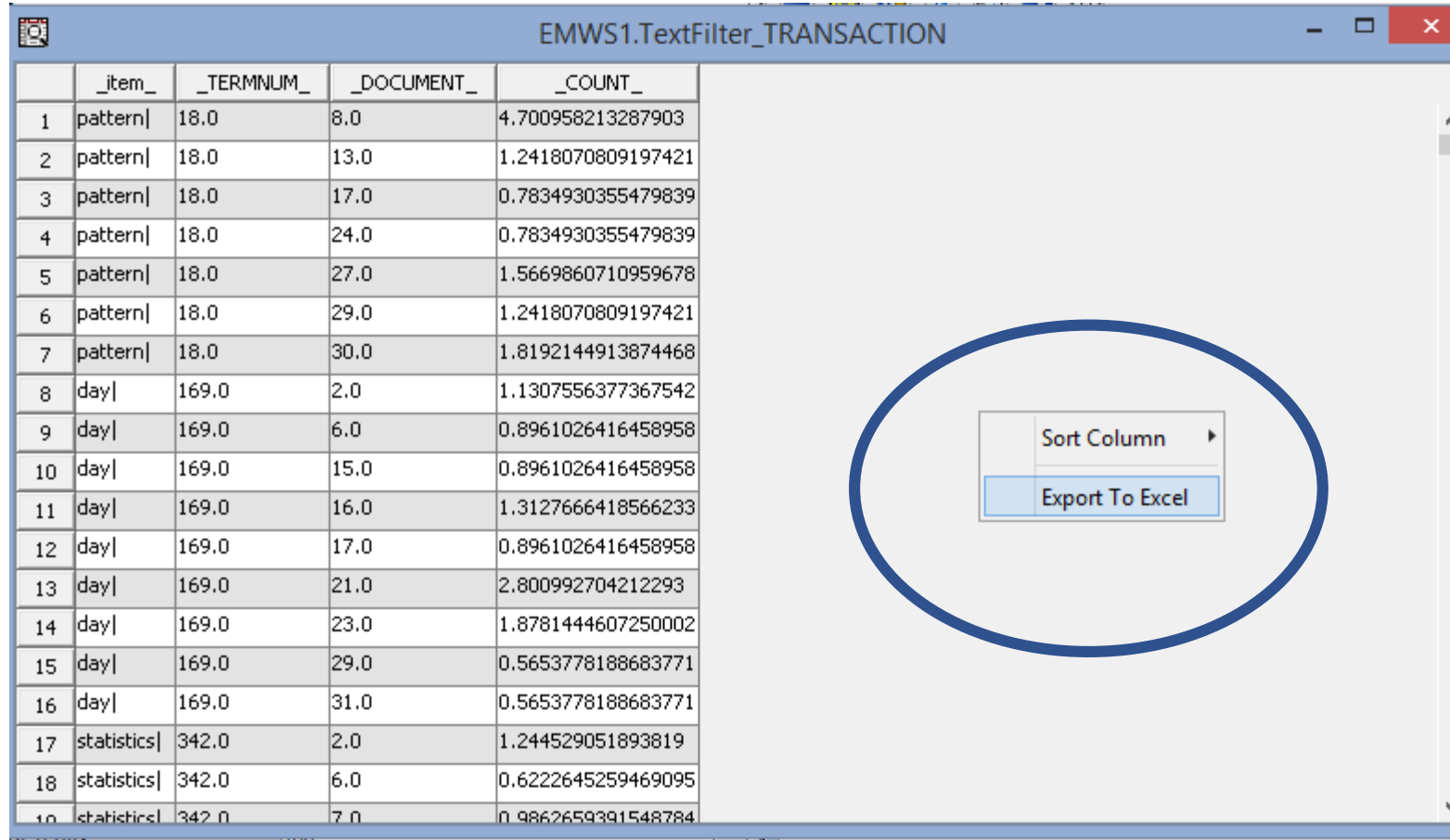
Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean
1	_item_		CLASS	0	.	.	.
2	_COUNT_		VAR	0	0.373201	7.325369	1.333
3	_DOCUMENT_		VAR	0	1	31	16.74
4	_TERMNUM_		VAR	0	18	10819	5776

EMWS1.TextFilter_TRANSACTION

Obs #	_item_	_TERMNUM_	_DOCUMENT_	_COUNT_
1	pattern	18	8	4.700958
2	pattern	18	13	1.241807
3	pattern	18	17	0.783493
4	pattern	18	24	0.783493
5	pattern	18	27	1.566986
6	pattern	18	29	1.241807
7	pattern	18	30	1.819214
8	day	169	2	1.130756
9	day	169	6	0.896103
10	day	169	15	0.896103
11	day	169	16	1.312767
12	day	169	17	0.896103
13	day	169	21	2.800993
14	day	169	23	1.878144
15	day	169	29	0.565218



If you click on “Browse” instead of “Explore” and right click in the term-by-document matrix, you can choose to export to Excel.



The screenshot shows a window titled "EMWS1.TextFilter_TRANSACTION" with a table of term-by-document data. The table has five columns: an index, a term (e.g., pattern, day, statistics), a term frequency (e.g., 18.0, 169.0, 342.0), a document frequency (e.g., 8.0, 2.0, 2.0), and a count (e.g., 4.700958213287903, 1.1307556377367542, 1.244529051893819). A right-click context menu is open over the table, with the "Export To Excel" option highlighted. The menu also includes a "Sort Column" option.

	item	_TERMNUM_	_DOCUMENT_	_COUNT_
1	pattern	18.0	8.0	4.700958213287903
2	pattern	18.0	13.0	1.2418070809197421
3	pattern	18.0	17.0	0.7834930355479839
4	pattern	18.0	24.0	0.7834930355479839
5	pattern	18.0	27.0	1.5669860710959678
6	pattern	18.0	29.0	1.2418070809197421
7	pattern	18.0	30.0	1.8192144913874468
8	day	169.0	2.0	1.1307556377367542
9	day	169.0	6.0	0.8961026416458958
10	day	169.0	15.0	0.8961026416458958
11	day	169.0	16.0	1.3127666418566233
12	day	169.0	17.0	0.8961026416458958
13	day	169.0	21.0	2.800992704212293
14	day	169.0	23.0	1.8781444607250002
15	day	169.0	29.0	0.5653778188683771
16	day	169.0	31.0	0.5653778188683771
17	statistics	342.0	2.0	1.244529051893819
18	statistics	342.0	6.0	0.6222645259469095
19	statistics	342.0	7.0	0.9862659391548784



Follow the steps outlined in the text below to evaluate the success of the import of the pdf files.

- Chakraborty, G., Pagolu, M., & Garla, S. (2014). Text mining and analysis: practical methods, examples, and case studies using SAS. SAS Institute.
<https://capella.skillport.com/skillportfe/main.action?path=summary/BOOKS/59026>
- Direct link to Chapter 5 – Data Transformation:
<https://capella.skillport.com/skillportfe/assetNonSSOLaunch.action?courseName= ss chapter:59026-350295958&courseType=7>

