1746-2014

Automatic Detection of Section Membership for SAS® Conference Paper Abstract Submissions: A Case Study

Dr. Goutam Chakraborty, Professor, Department of Marketing, Spears School of Business, Oklahoma State University
Murali Krishna Pagolu, Analytical Consultant, SAS® Institute Inc., Cary, NC

INTRODUCTION

Every year since its inception, SAS® and users of SAS® products have been actively making significant knowledge contributions to the SAS® user community using SAS® regional conferences and SAS® Global Forum as a platform. So far, thousands of papers were published in these conference proceedings under various topic sections. Each year, the number of contributions increase compared to its previous year and this trend is likely to continue in future. The array of SAS products, industry solutions and its user base is growing across the globe. You can anticipate a growing library of SAS® conference papers serving as free online education material showcasing several innovative applications of SAS® put into practice. What you see in the online proceedings are papers published after they got accepted through careful selection and scrutiny from a big pool of submissions every year.

Several scholars, efficient and experienced professionals from industry are handpicked, appointed as leaders/heads of individual sections based on their area of expertise, experience and knowledge. These leaders are bestowed with a huge task of reading through all the paper abstracts submitted and select those which qualify to be presented at the conference and later published in the conference proceedings. The number of paper abstracts submitted to SAS® Global Forum 2013 was rumored to be somewhere between 600 and 650. However, approximately 1/6th of them only were finally accepted. Paper acceptance criteria may depend on a lot of factors. Some of these factors are listed below:

- Type of submission (internal – submitted as a submission by SAS® employee or external – submitted by an external user of SAS®)
- Choice and relevance of the topic to the current section.
- Displaying theoretical accuracy and writing skills in the content.
- Showcasing a possible solution for a recurring problem in an industry or technology.
- Providing a business application using trending SAS® product(s) or technology.
- Discussing an innovative idea or technique.
- Preference pre-set by the section leaders and conference organizers in anticipation of attendees' background and interests.
- Range of competitive topics covered by other authors in their submissions for that section.

Though this list may not be exhaustive and accurate, one can determine that many of these factors play important roles in deciding the fate of an abstract submission. Except for the range of competitive abstracts submitted in that section by other authors, authors can make their best possible efforts to work on all other factors to increase the chance of their submission to get selected. Once authors have finished working on their abstract(s), the most important step that lies in their hands is to choose the appropriate section to submit their abstract. Some sections are so popular that they are often inundated with submissions creating a tough challenge for the evaluators to make their decisions in the selection process. Experienced authors may find it easy to narrow down to their top section choices (2 or 3) in which they may fit well according to the section description and the abstract topic. In such cases, their submission though rejected in one section may be accepted in other section due to one or more of the many reasons we discussed earlier for selection process. For example, a paper abstract discussing the usage of a unique segmentation method to distinctly identify several customer groups for better marketing and sales strategy may be applicable for both ‘Customer Intelligence’ and ‘Data Mining’ sections. A custom written SAS macro to address data integration issue may qualify for both ‘Data Integration’ and ‘Coder’s Corner’ sections. Hence, it is very critical for an author to determine the most appropriate choices of sections for submission to choose from a list of available sections.

OBJECTIVE

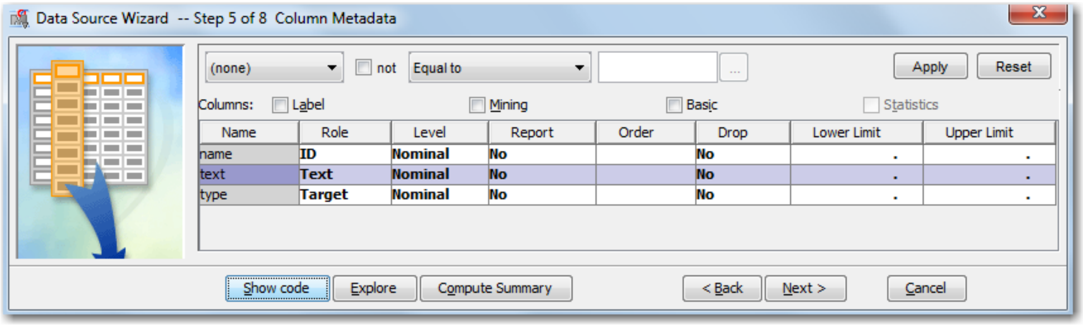
In this case study, we attempt to address the issue of determining the section membership of a paper abstract submission based on its content. For this purpose, we use SAS® Text Miner and SAS® Content Categorization Studio to develop rule based categorizer. This taxonomy should serve as an application to automatically score and identify the most relevant and appropriate conference section in which an abstract should be submitted for a better chance of acceptance. For this case study, we collected SAS® paper abstracts from SAS® Global Forum online proceedings http://support.sas.com/events/sasglobalforum/previous/online.html. We downloaded 466 papers from 2008 to 2012 encompassing 5 sections: ‘Business Intelligence’, ‘Reports’, ‘Data Mining’, ‘Statistical Analysis’ and ‘Systems Architecture’. Using %TMFILTER macro, we converted these papers in PDF file format to plain text files and parsed the content to retain only the abstracts in them. We have also created a SAS data set (**‘SGFpapers_sectionwise.sas7bdat’)** which holds these abstracts, file names and names of the sections to which they belong under three different columns. We use this data set in SAS® Text Miner to automatically build Boolean rules and use them in building rule based categorization models in SAS® Content Categorization Studio. In addition to this data set, we also created plain text files containing these abstracts in individual folders. These are used to test the categorization models in SAS® Content Categorization Studio. Data is available for download at the following URL: http://support.sas.com/publishing/bbu/zip/65646.zip.

STEP-BY-STEP INSTRUCTIONS

- Create a new project in SAS® Enterprise Miner and name it ‘**SGF_CS**’. Create a new diagram and name it ‘**Build_Rules**’. Create a library pointing to the location where the data resides using the project start code or File -> New -> Library menu.

- Add the SAS data set ‘**SGFpapers_bysection.sas7bdat**’ to the project and assign the roles of the variables as shown in Display 1. Variable ‘type’ should be assigned the role ‘Target’ and will be used to build a category prediction model using Text Builder feature in SAS® Text Miner.

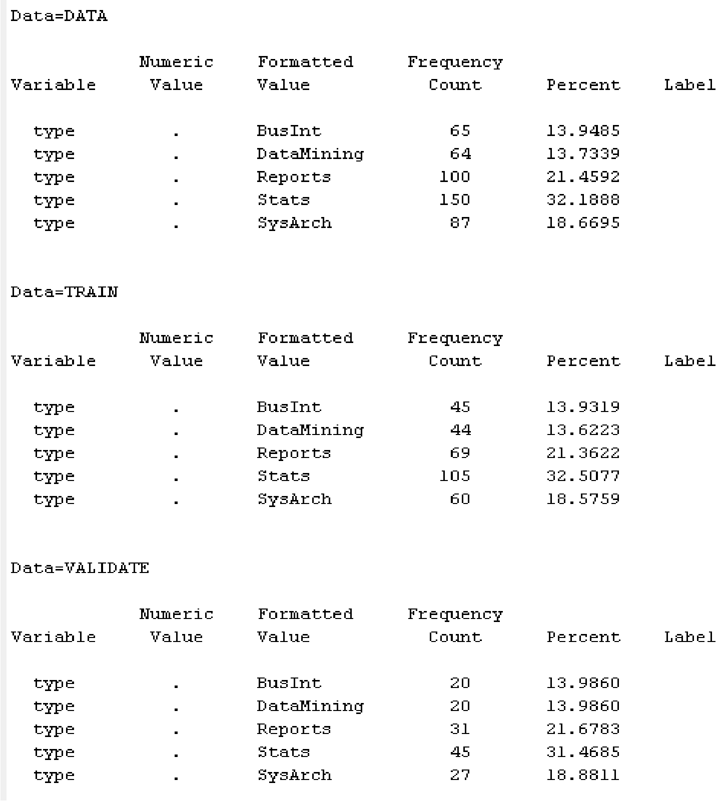
Display 1: Data Source Wizard for assigning roles to variables



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
name	ID	Nominal	No		No	.	.
text	Text	Nominal	No		No	.	.
type	Target	Nominal	No		No	.	.

- After the data set is added, drag the data source into the diagram as an 'Input Data' node. Now, connect a 'Data Partition' node and change the 'Data Set Allocations' property for Training, Validation and Test data sets to 70.0, 30.0 and 0.0 respectively. Run the data partition node to see the result as shown in Display 2. You will see that 'Stats' category contains more number of abstracts compared to other sections. This is not intentional but in general more papers are published in 'Statistical Analysis' section every year hence the difference in counts.

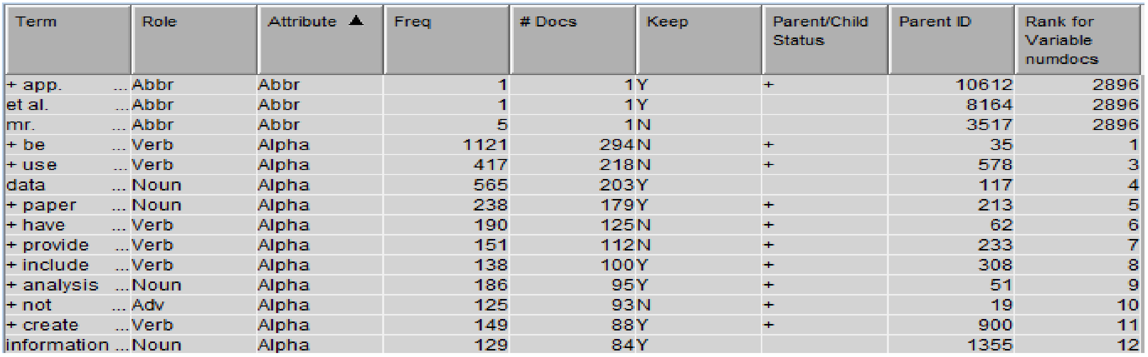
Display 2: Distribution of abstracts by section name in Train, Validation and Test groups



```
Data=DATA

            Numeric    Formatted     Frequency
Variable      Value    Value             Count      Percent    Label

  type          .      BusInt               65      13.9485
  type          .      DataMining           64      13.7339
  type          .      Reports             100      21.4592
  type          .      Stats               150      32.1888
  type          .      SysArch              87      18.6695


Data=TRAIN

            Numeric    Formatted     Frequency
Variable      Value    Value             Count      Percent    Label

  type          .      BusInt               45      13.9319
  type          .      DataMining           44      13.6223
  type          .      Reports              69      21.3622
  type          .      Stats               105      32.5077
  type          .      SysArch              60      18.5759


Data=VALIDATE

            Numeric    Formatted     Frequency
Variable      Value    Value             Count      Percent    Label

  type          .      BusInt               20      13.9860
  type          .      DataMining           20      13.9860
  type          .      Reports              31      21.6783
  type          .      Stats                45      31.4685
  type          .      SysArch              27      18.8811
```

- Connect a 'Text Parsing' node to the data partition node and run it with the default property settings. Display 3 shows a partial screenshot of the text parsing node results displaying list of terms found after the abstracts are parsed. You can see the terms with Attribute type 'Abbr' (Abbreviation) are not so frequently occurring in the data set. Also, there are generic terms such as data, paper, include, information etc. with the 'keep' status – 'Y' in the terms list. It means these terms are kept or retained in the process flow for the next node/feature to use. Terms with 'keep' status – 'N' in the terms list are thus excluded from further analysis. Similarly, terms such as 'miner', 'enterprise miner' though represent the same thing but appear as different terms in the abstracts. Hence, synonyms should be added to the list whenever possible to reduce the size of terms list.

Display 3: Partial output of Text parsing results showing term list



Term	Role	Attribute ▲	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ app. ...	Abbr	Abbr	1	1	Y	+	10612	2896
et al. ...	Abbr	Abbr	1	1	Y		8164	2896
mr. ...	Abbr	Abbr	5	1	N		3517	2896
+ be ...	Verb	Alpha	1121	294	N	+	35	1
+ use ...	Verb	Alpha	417	218	N	+	578	3
data ...	Noun	Alpha	565	203	Y		117	4
+ paper ...	Noun	Alpha	238	179	Y	+	213	5
+ have ...	Verb	Alpha	190	125	N	+	62	6
+ provide ...	Verb	Alpha	151	112	N	+	233	7
+ include ...	Verb	Alpha	138	100	Y	+	308	8
+ analysis ...	Noun	Alpha	186	95	Y	+	51	9
+ not ...	Adv	Alpha	125	93	N	+	19	10
+ create ...	Verb	Alpha	149	88	Y	+	900	11
information ...	Noun	Alpha	129	84	Y		1355	12

- Based on our analysis on text parsing results in the previous step, you can make the following changes to text parsing node properties and re-run the node.

  - Add 'Abbr' and 'Num' to the 'Ignore Parts of Speech' property
  - Add 'Abbr' to 'Ignore Types of Attributes' property

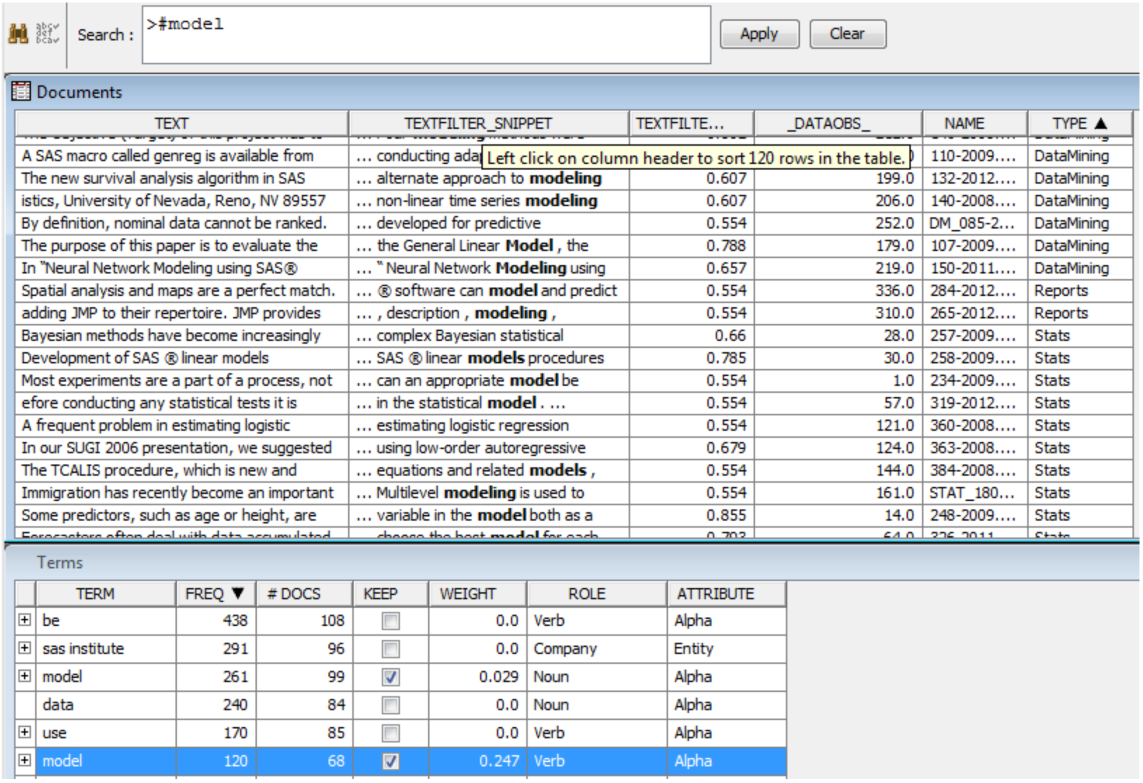
- Connect a 'Text Filter' node to the text parsing node. In the text filter node property panel, change the 'Term Weight' weightings property to 'Mutual Information' and run the node. The categorical variable is defined with a role of 'Target' in the data source; hence, this is the most appropriate term weight to use. Let the other properties set to default.

- Now that the text filtering node is run at least once, click on the ellipsis button next to 'Filter Viewer' under 'Results' property. It opens an 'Interactive Filter Viewer' providing the list of terms, total frequency of occurrence in the corpus, number of documents in which they occur at least once, keep flag, term weight, role and attribute.

- As discussed in the previous steps, you may choose to modify the keep flag to either drop or add certain terms based on your intuition and frequency of occurrence of terms to arrive at a better classification model for the section category. If there are terms which need to be closely investigated, you may choose to use that term to search and find the abstracts data containing that term. For example, the term 'model' can mean either a statistical model or a predictive model. Hence, right click on that term in the Terms list and click 'Add Term to Search Expression'.

- Click **Apply** next to the Search window to find all the documents containing this term. Display 4 shows the document search result for the term 'model'. All the words 'modeling', 'models', 'modeled' stemmed to the root word 'model' can be found highlighted in bold from the 'TESTFILTER_SNIPPET' column. You can also see the 'Type' of the abstract to which these documents belong in the same table. You will see that the documents containing the term 'model' are fairly distributed between the topic sections 'Data Mining' and 'Stats'.

Display 4: Searching for terms in documents using Interactive Filter Viewer



Search : >#model

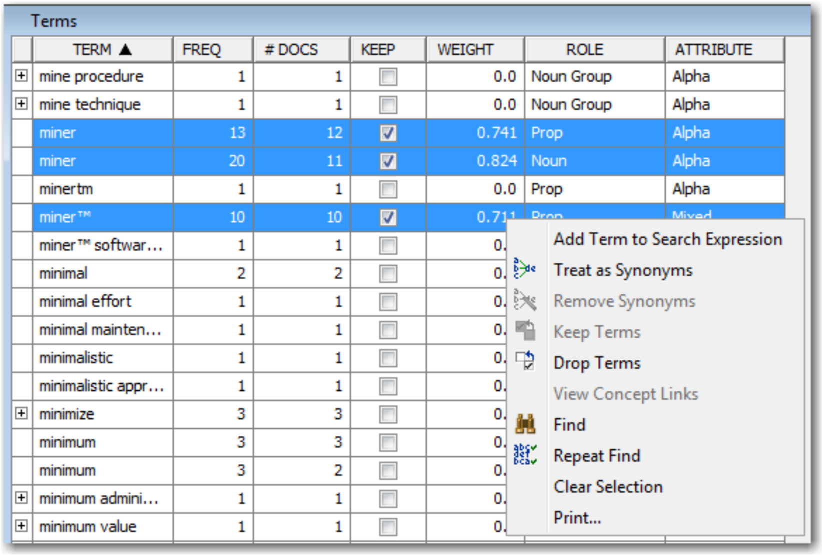
TEXT	TEXTFILTER_SNIPPET	TEXTFILTE…	_DATAOBS_	NAME	TYPE ▲
A SAS macro called genreg is available from	… conducting ada Left click on column header to sort 120 rows in the table.			110-2009.…	DataMining
The new survival analysis algorithm in SAS	… alternate approach to **modeling**	0.607	199.0	132-2012.…	DataMining
istics, University of Nevada, Reno, NV 89557	… non-linear time series **modeling**	0.607	206.0	140-2008.…	DataMining
By definition, nominal data cannot be ranked.	… developed for predictive	0.554	252.0	DM_085-2…	DataMining
The purpose of this paper is to evaluate the	… the General Linear **Model** , the	0.788	179.0	107-2009.…	DataMining
In "Neural Network Modeling using SAS®	… " Neural Network **Modeling** using	0.657	219.0	150-2011.…	DataMining
Spatial analysis and maps are a perfect match.	… ® software can **model** and predict	0.554	336.0	284-2012.…	Reports
adding JMP to their repertoire. JMP provides	… , description , **modeling** ,	0.554	310.0	265-2012.…	Reports
Bayesian methods have become increasingly	… complex Bayesian statistical	0.66	28.0	257-2009.…	Stats
Development of SAS ® linear models	… SAS ® linear **models** procedures	0.785	30.0	258-2009.…	Stats
Most experiments are a part of a process, not	… can an appropriate **model** be	0.554	1.0	234-2009.…	Stats
efore conducting any statistical tests it is	… in the statistical **model** . …	0.554	57.0	319-2012.…	Stats
A frequent problem in estimating logistic	… estimating logistic regression	0.554	121.0	360-2008.…	Stats
In our SUGI 2006 presentation, we suggested	… using low-order autoregressive	0.679	124.0	363-2008.…	Stats
The TCALIS procedure, which is new and	… equations and related **models** ,	0.554	144.0	384-2008.…	Stats
Immigration has recently become an important	… Multilevel **modeling** is used to	0.554	161.0	STAT_180…	Stats
Some predictors, such as age or height, are	… variable in the **model** both as a	0.855	14.0	248-2009.…	Stats

Terms

	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
⊞	be	438	108	□	0.0	Verb	Alpha
⊞	sas institute	291	96	□	0.0	Company	Entity
⊞	model	261	99	☑	0.029	Noun	Alpha
	data	240	84	□	0.0	Noun	Alpha
⊞	use	170	85	□	0.0	Verb	Alpha
⊞	model	120	68	☑	0.247	Verb	Alpha

- In this case study, you do not require a great deal of modification to the terms list. Optionally, you can start creating a synonym list based on the closely related terms. You can highlight those terms which represent the same thing, right click and select 'Treat as Synonyms'. For example, terms 'miner' and 'miner^{TM}' can be treated as synonyms for SAS Enterprise Miner (Display 5).

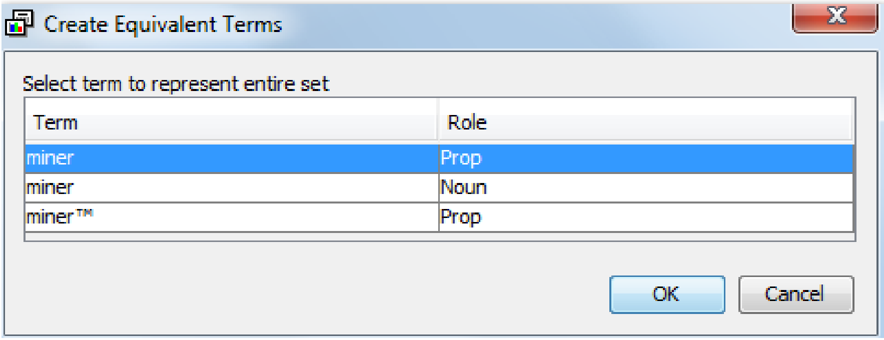
Display 5: Build synonyms list choosing terms meaning the same



	TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
⊞	mine procedure	1	1	□	0.0	Noun Group	Alpha
⊞	mine technique	1	1	□	0.0	Noun Group	Alpha
	miner	13	12	☑	0.741	Prop	Alpha
	miner	20	11	☑	0.824	Noun	Alpha
	minertm	1	1	□	0.0	Prop	Alpha
	miner™	10	10	☑	0.711	Prop	Mixed
	miner™ softwar…	1	1	□	0.		
	minimal	2	2	□	0.		
	minimal effort	1	1	□	0.		
	minimal mainten…	1	1	□	0.		
	minimalistic	1	1	□	0.		
	minimalistic appr…	1	1	□	0.		
⊞	minimize	3	3	□	0.		
	minimum	3	3	□	0.		
	minimum	3	2	□	0.		
⊞	minimum admini…	1	1	□	0.		
⊞	minimum value	1	1	□	0.		

- Choose one of these highlighted terms to be used as the equivalent term to represent all of these synonymous terms in the next pop-up window (Display 6). Similarly, in this case study you may also treat terms such as 'mine', 'data mine', 'data mining', 'mining' as synonyms with 'mining' as the equivalent term representing all these terms. Hence, in that case, you can select all these terms at once and use the 'Treat as Synonyms' option to create the synonym list. It is important to export the synonyms list you have created by clicking on File--> Export Synonyms. Give a name for the data set and store it in the library you have created from the project start up code (default). Close the 'Interactive Filter Viewer' window and click 'Ok' on the prompt window to save results.

Display 6: Choose the term to represent the entire data set



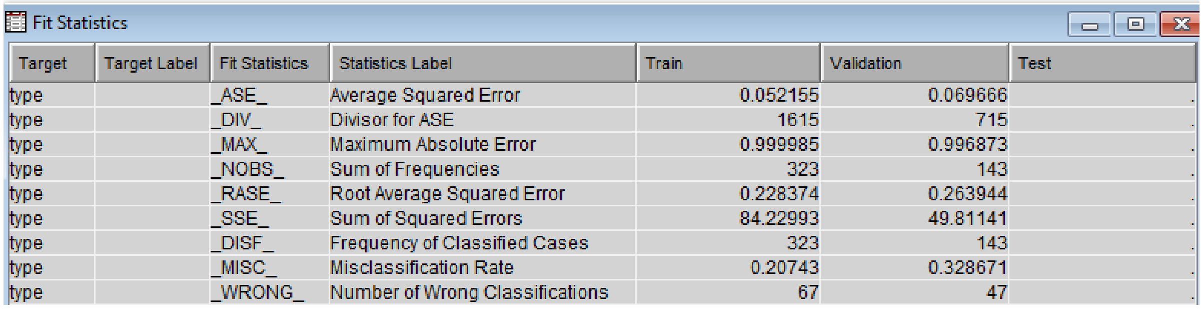
Create Equivalent Terms

Select term to represent entire set

Term	Role
miner	Prop
miner	Noun
miner™	Prop

- Drag a 'Text Rule builder' node into the diagram, connect it to the text filtering node and run using the default properties. Once the node run is complete, click on Results to view the output. In the Fit Statistics, you will find the misclassification rate to be approximately 20% for the training data and 32% for the validation data (Display 7). This is a very good model given that SAS® Text Miner has automatically built rules to classify abstracts using the training corpus. Close the results window.

Display 7: Fit Statistics results from Text Rule Builder node

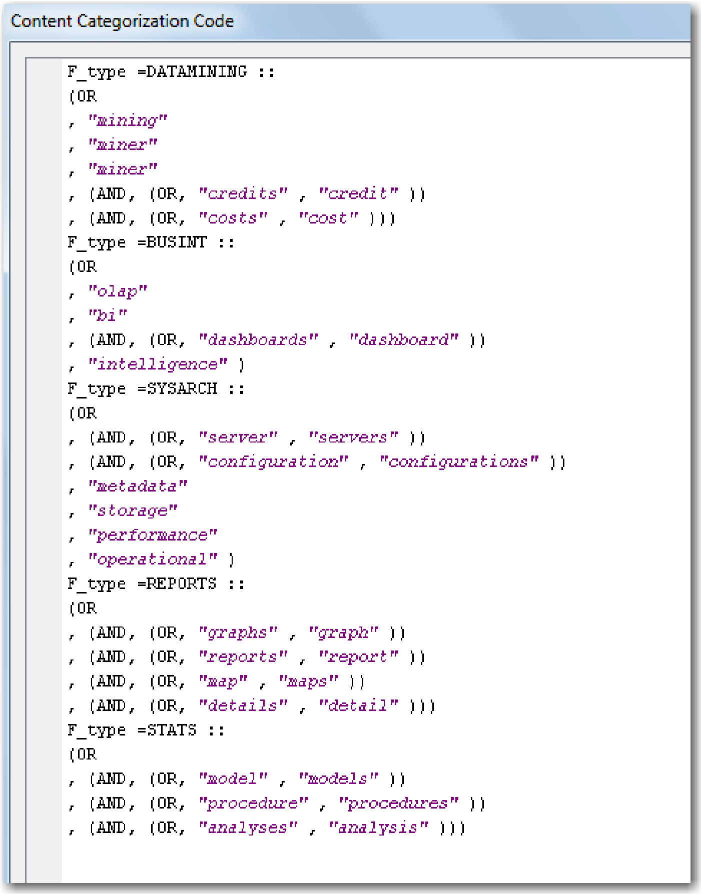


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
type		_ASE_	Average Squared Error	0.052155	0.069666	.
type		_DIV_	Divisor for ASE	1615	715	.
type		_MAX_	Maximum Absolute Error	0.999985	0.996873	.
type		_NOBS_	Sum of Frequencies	323	143	.
type		_RASE_	Root Average Squared Error	0.228374	0.263944	.
type		_SSE_	Sum of Squared Errors	84.22993	49.81141	.
type		_DISF_	Frequency of Classified Cases	323	143	.
type		_MISC_	Misclassification Rate	0.20743	0.328671	.
type		_WRONG_	Number of Wrong Classifications	67	47	.

- If you click on the ellipsis button next to 'Content Categorization Code' under the 'Score' property, you will find the rule expressions automatically built by the text rule builder node (Display 8). These rules are in the same syntax as that of 'SAS Content Categorization Studio' and hence can be used for building a Boolean rule based categorizer for all those section categories.

Display 8: Automatic Content Categorization Code generated by text rule builder node

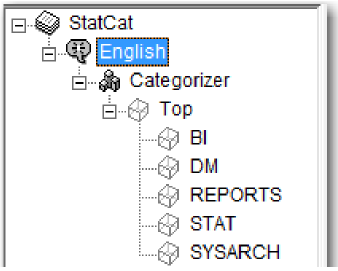


```
F_type =DATAMINING ::
(OR
, "mining"
, "miner"
, "miner"
, (AND, (OR, "credits" , "credit" ))
, (AND, (OR, "costs" , "cost" )))
F_type =BUSINT ::
(OR
, "olap"
, "bi"
, (AND, (OR, "dashboards" , "dashboard" ))
, "intelligence" )
F_type =SYSARCH ::
(OR
, (AND, (OR, "server" , "servers" ))
, (AND, (OR, "configuration" , "configurations" ))
, "metadata"
, "storage"
, "performance"
, "operational" )
F_type =REPORTS ::
(OR
, (AND, (OR, "graphs" , "graph" ))
, (AND, (OR, "reports" , "report" ))
, (AND, (OR, "map" , "maps" ))
, (AND, (OR, "details" , "detail" )))
F_type =STATS ::
(OR
, (AND, (OR, "model" , "models" ))
, (AND, (OR, "procedure" , "procedures" ))
, (AND, (OR, "analyses" , "analysis" )))
```

- Launch SAS® Content Categorization Studio, create a new project and name it SGF_Cat_CS_2. Right click on 'SGF_Cat_CS_2' and click 'Add Language'. Select 'English' as the language and click Ok.

- Right click on 'English' and select the option 'Create Categorizer from Directories'. Browse to the location on the PC where the paper abstracts in raw text file format are stored separated by section. Navigate to the 'Top' subfolder contained within the **'SGF_SECTIONWISE**' folder and click Ok to create the categories based on the folder structure (Display 9). Categories are created as BI – Business Intelligence, DM – Data Mining, REPORTS – Visualization and Reporting, STAT – Statistical Analysis and SYSARCH – Systems Architecture.

Display 9: Categories created from an existing folder structure



- Select any of these section categories and click on ‘Data’ tab. You will find the training path automatically populated for each of these categories since you created them using existing folder structure instead of creating them manually.

- Change the Training and Testing Paths of the categories and point them to the designated ‘Train’ and ‘Test’ folders (Table 1) to prepare for building a Statistical Categorization model.

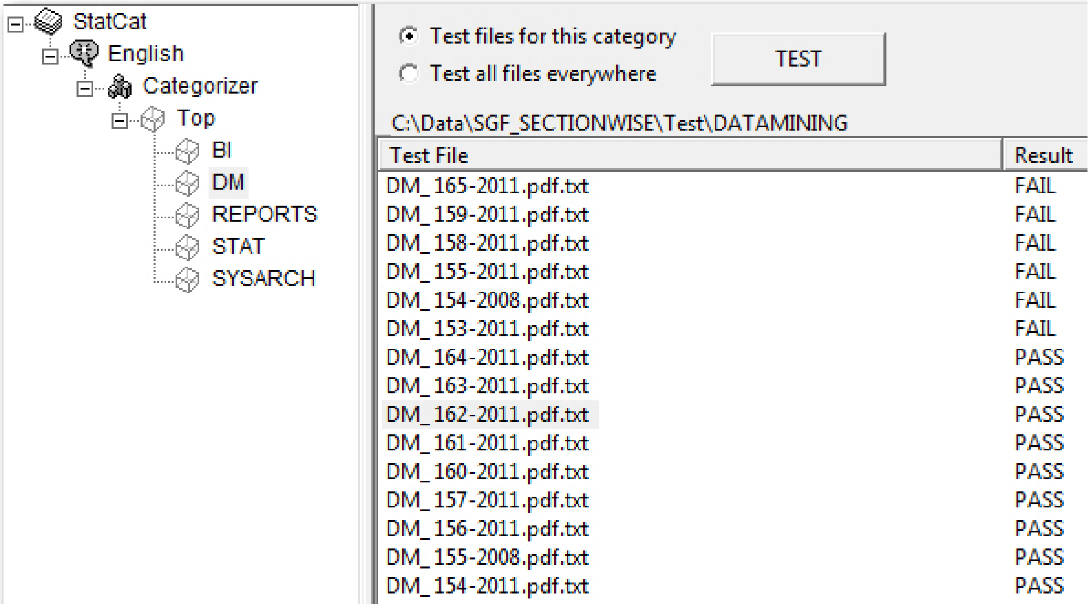
Table 1: Testing and Training Paths by section category for statistical model

Category	Training Path	Testing Path
BI	C:\Data\SGF_SECTIONWISE\Train\BUSINT	C:\Data\SGF_SECTIONWISE\Test\BUSINT
DM	C:\Data\SGF_SECTIONWISE\Train\DATAMINING	C:\Data\SGF_SECTIONWISE\Test\DATAMINING
REPORTS	C:\Data\SGF_SECTIONWISE\Train\REPORTS	C:\Data\SGF_SECTIONWISE\Test\REPORTS
STAT	C:\Data\SGF_SECTIONWISE\Train\STAT	C:\Data\SGF_SECTIONWISE\Test\STAT
SYSARCH	C:\Data\SGF_SECTIONWISE\Train\SYSARCH	C:\Data\SGF_SECTIONWISE\Test\SYSARCH

- Now that the training and testing paths are set for all the categories, click on Build -- > Build Statistical Categorizer to generate a statistical model. Once you receive a message ‘Build Successful’, click Ok and go to the Testing tab on any of the categories, for example DM (Data Mining). You will find a list of files populated from the ‘Test’ folder of the category ready to be tested against the statistical model you just built.

- Click ‘Test’ and view the results to find out how many of those files have failed the test and how many passed the test (Display 10). As you can observe, there are a few files which failed the test but there are some which passed. You may double-click on any of the listed files to open the actual abstract contained within the file. However, statistical categorizer is a black box model which is why you cannot see the rules working behind the scenes for categorization process. There is not much you can do to better the performance of a statistical model other than increasing the size of training corpus for each of these categories. Statistical models largely depend on the quality of training documents by which they are truly separated by each category with respect to another category.

Display 10: Test results for DM (Data Mining) category using statistical model



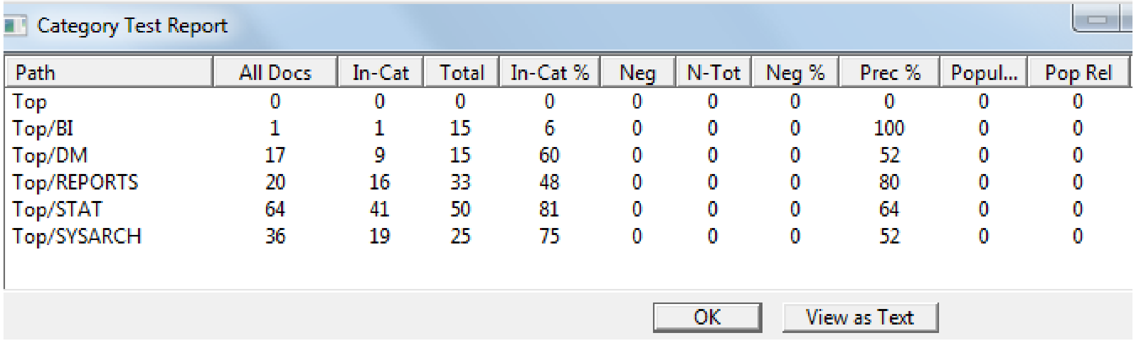
● Test files for this category
○ Test all files everywhere

C:\Data\SGF_SECTIONWISE\Test\DATAMINING

Test File	Result
DM_165-2011.pdf.txt	FAIL
DM_159-2011.pdf.txt	FAIL
DM_158-2011.pdf.txt	FAIL
DM_155-2011.pdf.txt	FAIL
DM_154-2008.pdf.txt	FAIL
DM_153-2011.pdf.txt	FAIL
DM_164-2011.pdf.txt	PASS
DM_163-2011.pdf.txt	PASS
DM_162-2011.pdf.txt	PASS
DM_161-2011.pdf.txt	PASS
DM_160-2011.pdf.txt	PASS
DM_157-2011.pdf.txt	PASS
DM_156-2011.pdf.txt	PASS
DM_155-2008.pdf.txt	PASS
DM_154-2011.pdf.txt	PASS

- Click ‘Testing -- > Full Test Report’ to generate precision and recall scores specific to each category (Display 11). If you look at the recall values (In-Cat% column), you can clearly observe a very low score (6%) for BI, medium score (48%) for REPORTS, reasonable scores for DM, SYSARCH and very good score (81%) for STAT categories. This is a basic model based on the statistical analysis of training corpus that you can build very quickly using Content Categorization Studio.

Display 11: Full Test Report results of all categories using Statistical model



Category Test Report

Path	All Docs	In-Cat	Total	In-Cat %	Neg	N-Tot	Neg %	Prec %	Popul...	Pop Rel
Top	0	0	0	0	0	0	0	0	0	0
Top/BI	1	1	15	6	0	0	0	100	0	0
Top/DM	17	9	15	60	0	0	0	52	0	0
Top/REPORTS	20	16	33	48	0	0	0	80	0	0
Top/STAT	64	41	50	81	0	0	0	64	0	0
Top/SYSARCH	36	19	25	75	0	0	0	52	0	0

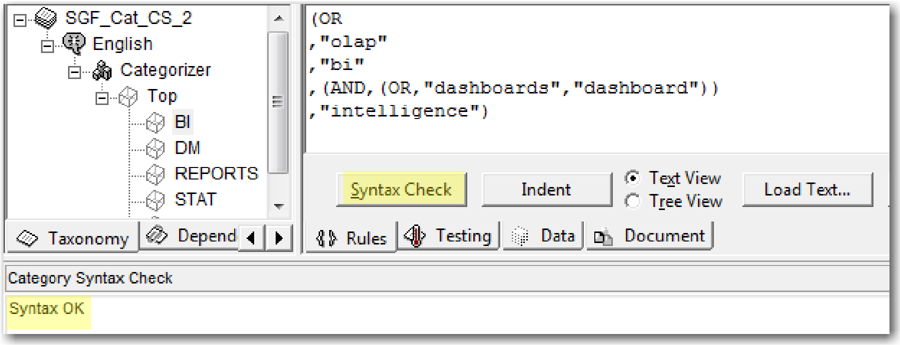
- Now you have a base model (statistical categorizer model) in Content Categorization Studio to compare against the Boolean rule based model you can build using content categorization code automatically generated from text rule builder node of SAS® Text Miner. As you know, you do not require setting training data to build rule based categorization models. Hence, you may now change your testing paths for all the categories as shown in Table 2 and keep the training paths blank.

Table 2: Testing Paths by section category for rule based model

Category	Testing Path
BI	C:\Data\SGF_SECTIONWISE\Top\BUSINT
DM	C:\Data\SGF_SECTIONWISE\Top\DATAMINING
REPORTS	C:\Data\SGF_SECTIONWISE\Top\REPORTS
STAT	C:\Data\SGF_SECTIONWISE\Top\STAT
SYSARCH	C:\Data\SGF_SECTIONWISE\Top\SYSARCH

- Copy automatically generated content categorization code that you have previously generated using text rule builder node and paste them under ‘Rules’ tab for each of the categories. Click on ‘Syntax Check’ button each time you copy and paste those rules for every category (Display 12).

Display 12: Syntax check of content categorization code in Rules tab



```
(OR
,"olap"
,"bi"
,(AND,(OR,"dashboards","dashboard"))
,"intelligence")
```

● Text View
○ Tree View

Category Syntax Check

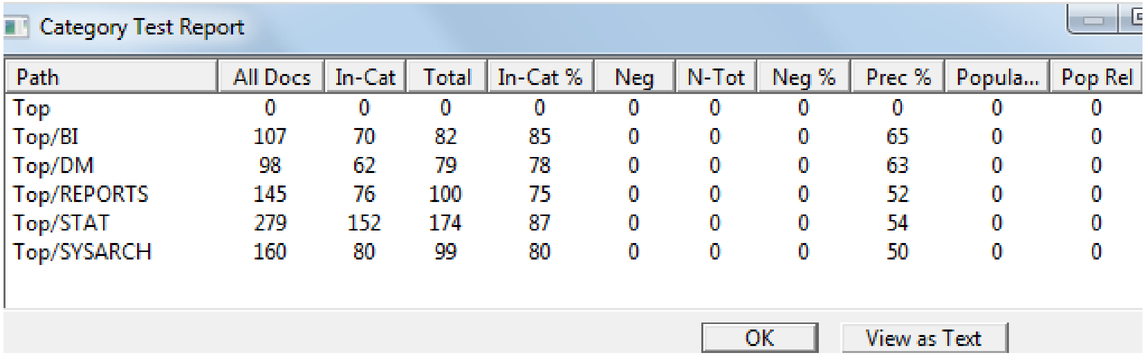
Syntax OK

- Go to Build -- > Build Rulebased Categorizer to build a Boolean rule based categorization model using those rule expressions that you have imported from SAS® Text Miner. You will receive a message 'Build Successful' once you were able to successfully build a Boolean Rule based Categorization model.

- Since you have set the testing paths for each of the 5 categories, you may click on any category and go to 'Testing' tab to view the test files. Click 'Test' to test the files based on the Boolean Rulebased categorization model you have built. You will find the test results (pass or fail) and relevancy scores for each of the test file that passed the test.

- Click 'Testing -- > Full Test Report' to generate a full test report on the model performance with recall and precision scores (Display 13). In general, you will observe that more files pass test in this model compared to the statistical model you have built previously. This is because Boolean rule based models are flexible to write your own rules based on linguistic terms and incorporate Boolean operators for improved accuracy. In this case, you have just exported the automatic rules generated from Text Rule Builder node into SAS® Content Categorization Studio and used them 'as is'. However, after careful examination of the test documents and using the domain knowledge of individual section categories these rules can be further modified to improve accuracy.

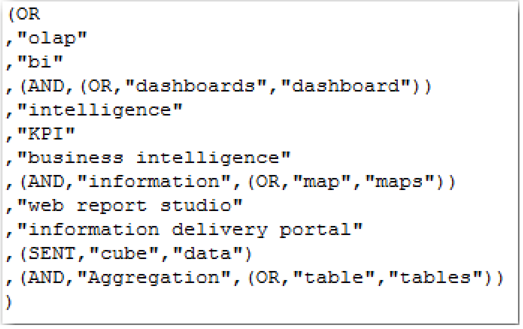
Display 13: Full Test Report results of all categories using Boolean Rule based model



Path	All Docs	In-Cat	Total	In-Cat %	Neg	N-Tot	Neg %	Prec %	Popula...	Pop Rel
Top	0	0	0	0	0	0	0	0	0	0
Top/BI	107	70	82	85	0	0	0	65	0	0
Top/DM	98	62	79	78	0	0	0	63	0	0
Top/REPORTS	145	76	100	75	0	0	0	52	0	0
Top/STAT	279	152	174	87	0	0	0	54	0	0
Top/SYSARCH	160	80	99	80	0	0	0	50	0	0

- Click on 'BI' category and go to the testing tab and carefully examine the files which have failed the test. You will find many terms which are unique to this section category that were not picked up by the text rule builder node during the automatic rule generation process. Terms such as 'information map(s)', 'web report studio', 'information delivery portal' and 'KPI' can specifically identify the topic "business intelligence (BI)" because these are the names of products and features used in SAS Enterprise Business Intelligence suite. Whenever these are identified in the documents, you can conveniently relate them to the BI category. Modify the rules in this category as shown in the Display 14. As you can observe, terms such as cube, data, aggregation and table(s) are also added to the rules bound by Boolean operators to ensure more variety of patterns captured.

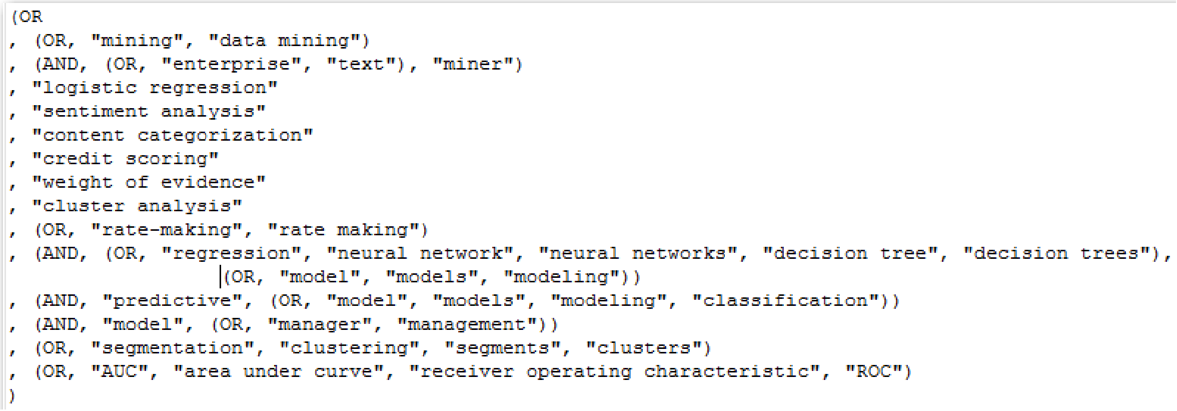
Display 14: Modified Boolean rules for Business Intelligence (BI) category



```
(OR
,"olap"
,"bi"
,(AND,(OR,"dashboards","dashboard"))
,"intelligence"
,"KPI"
,"business intelligence"
,(AND,"information",(OR,"map","maps"))
,"web report studio"
,"information delivery portal"
,(SENT,"cube","data")
,(AND,"Aggregation",(OR,"table","tables"))
)
```

- Similarly, rules can be modified for the data mining category to include terms specifically related to predictive modeling (regression, decision trees and neural network), clustering, model comparison (receiver operating characteristic) and so on. Terms representing products or features such as enterprise miner, credit scoring, model manager etc. related to data mining field are generally found useful in modifying the rules to match the category (Display 15). It is also important to remember that this rule modification is an iterative process requiring careful understanding of terms that can lead to category matching.

Display 15: Modified Boolean rules for Data Mining (DM) category

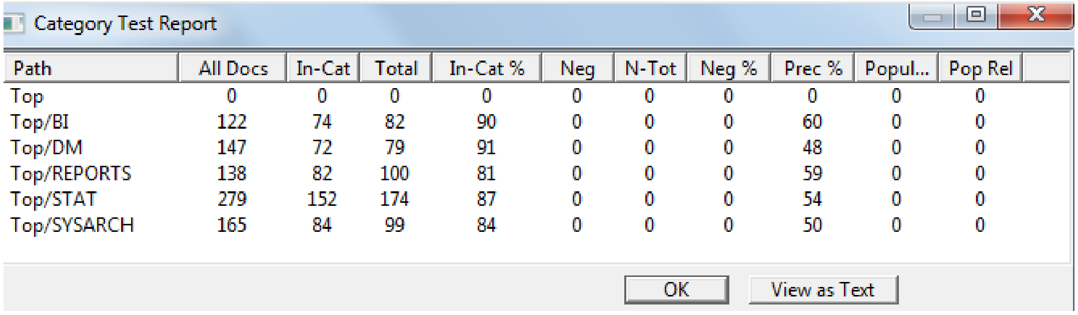


```
(OR
, (OR, "mining", "data mining")
, (AND, (OR, "enterprise", "text"), "miner")
, "logistic regression"
, "sentiment analysis"
, "content categorization"
, "credit scoring"
, "weight of evidence"
, "cluster analysis"
, (OR, "rate-making", "rate making")
, (AND, (OR, "regression", "neural network", "neural networks", "decision tree", "decision trees"),
                  (OR, "model", "models", "modeling"))
, (AND, "predictive", (OR, "model", "models", "modeling", "classification"))
, (AND, "model", (OR, "manager", "management"))
, (OR, "segmentation", "clustering", "segments", "clusters")
, (OR, "AUC", "area under curve", "receiver operating characteristic", "ROC")
)
```

- You may continue to analyze the terms which may represent the products, features or capabilities that better define a particular category and test them well before moving on to the next category. Once all the category rules are modified, rebuild the Boolean rule based categorizer model again and generate the Full test report (See Display 16) to compare its performance against other models you have built so far. You will observe that the model accuracy has increased overall compared to using either the default automatic rules generated from text rule builder node in SAS® Text Miner or the Statistical categorizer model.

- It is important to remember that usually categorization models are not 100% in their predictive ability. Hence, even if you write rules of high precision and quality it can only improve the performance to a certain extent after which it may degrade with the addition of more terms and/or rules there by losing its generality. Hence, we suggest you to practice rule writing and ensuring that those rules are neither too broad nor too specific. This is a very subjective job and the style, approach of modifying these rules can vary from analyst to analyst.

Display 16: Full Test Report using Boolean rule based model with modified rules



Path	All Docs	In-Cat	Total	In-Cat %	Neg	N-Tot	Neg %	Prec %	Popul...	Pop Rel
Top	0	0	0	0	0	0	0	0	0	0
Top/BI	122	74	82	90	0	0	0	60	0	0
Top/DM	147	72	79	91	0	0	0	48	0	0
Top/REPORTS	138	82	100	81	0	0	0	59	0	0
Top/STAT	279	152	174	87	0	0	0	54	0	0
Top/SYSARCH	165	84	99	84	0	0	0	50	0	0

SUMMARY

- SAS® Content Categorization Studio is an easy-to-use point and click interface used in quickly building models for automatic text categorization process.
- Statistical categorizer utilizes a set of documents from each category in the taxonomy to train the model. However, in terms of model performance statistical categorizer often performs below par.
- Boolean rule based categorizer works well when the rule terms and Boolean operators are carefully chosen to categorize documents. You can iteratively build the model while testing the rules on a set of documents. It has an additional advantage that you don't need a separate set of documents to train the model.
- Text rule builder node in SAS® Text Miner is a powerful feature useful to generate preliminary Boolean rule expressions which can be exported to SAS® Content Categorization Studio. It requires a set of documents separated by category to train the model and generate rules.

REFERENCES

SAS® Enterprise Miner 12.1: Reference Help Documentation. Cary, NC.

Reprinted with permission of SAS Institute Inc. from *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. Copyright 2013. SAS Institute Inc.

TRADEMARKS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

CONTACT INFORMATION

The authors welcome and encourage any questions, feedback, remarks, both on- and off-topic via email.

Goutam Chakraborty, Oklahoma State University, Stillwater, OK, goutam.chakraborty@okstate.edu

Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate program at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS®.

Murali Krishna Pagolu, SAS Institute Inc., Cary, NC, murali.pagolu@sas.com

Murali Pagolu is a Business Analytics Consultant at SAS Institute Inc. He has over 4 years of experience using SAS® software focused on Database Marketing, Marketing Research, Data mining, Text Mining and CRM Applications.