Here are steps to create a SAS Content Categorization project. There will be steps to create both a category rule as well as a concept rule. Each rule will be built or compiled, and then testing on the rule will be done on a set of documents that is part of the zip file.

Part 1: Creating a project

1. Create New Project and set the Location where the new Project Folder will be stored.



Now we will begin creating rules in the newly created *Demo* project.

2. Add language to the new project Demo

re tos tem oue reput vegey vege tong vegen temp vegen temp temp temp temp temp temp temp temp	1997 - 19
	6
Add Language	(C
Espand Fully	and the second s
	9
	100
	U
	S
	03
	5
	W
	1 10 40
	12:34
Cancerny @ Depend () 49 Rules @ Testing Data by Document	Thursday
Redy NUM	

DIMO:82 - 545 Content Categorization Studio	0 9 2	
File Edit View Build Project Category Concept Testing Document Server Help		
		e
		٥
		9
		0
		E.
		S
		0
Select a Language		
Fredick English-UTF8		w
Reference UTIP		
Spanish-UTF8		
		2
S Tasonomy S Depend ()) (3) Rules (⊕ Tetting) □ Data (B) Document.	NUM	 ↓ ↓> 12:34 Thursday 14-03-2013



Part 2: Creating a Category

1. To create a category taxonomy, right click on English and choose Enable Categorizer.



DEMO.tk2 - SAS Content Categorization Studio	Man, C. Andreas, Manufrid		-
Eile Edit View Build Project Category Cogcept Testing Document Server Help			1
D 📽 🖬 🕼 🎯 🥥 🌲 🗛 😘 🕾 🤔 🥐 🧣			
B S DEMO			e
E- op Camponer			
			0
			0
			0
			-
			15
			S
			0
			5
			W
Taxonomy Depend () 4 Rules Testing Data Data Do Document		14 March 2013 Thurddy	13:12 Thursday 4-03-2013
			_

2. Next, create a category by right mouse clicking on Top and select Add Category

DEMO.152 - SAS Content Categorization Studio	- 0 ×	
Eile Edit View Build Poject Satepoy Concept Testing Document Server Help		1 1 1 1
		6
i ⊕ English		e
Add Category		
Paste Paste Seniole Norde		
Paste Symbolic Link		
Paste as Macro		9
Unterlate successforms		
		1-
		S
		0
		W
		123
		F
		1318
C Tassnew & Depend () 43 Rules & Testing U Data (b) Document		Thursday
Tendo Internet	NUM	14-03-2013

CEMO tb2 - SAS Content Categorization Studio		
Elia Edit Yiew Build Broject Category Cegocept Testing Document Server Help		
E ŵ DENO E ∰ English		e
B de Camporten B - O Top - O Bounesal		
		O
		5
		0
		- <u>15</u>
		S
		0
		5
		W
		E/
		1000
Synta: Check Indext G Text View Load Text. Expand Forms Server Query		• 🕽 👀
Tannemy & Depend () 3 8 Bules & Tetting U Data (3) Document		Thursday
Rody	Ln 1 NUM	14-03-2013

On the rule tab of the Business category write the rules using Text View or Tree View.

3. Within Text View, in the rule tab insert the following rule:

(SENT, business, (OR, profit, profitable, break even))

DEMO.tk2 - SAS Content Categorization Studio	Ballander Bauffer		
Eile Edit View Build Project Category C	ogcept Testing Document Server Help		
D 🗃 🖬 🙆 🎯 🥥 🌲 🍬 😁 🕄			
E DEMO (SENT, "E	usiness", (08, "profit", "profitable", "break even"))		e
😑 🎝 Categorizer			-
Business			
			٥
			9
			0
			L.
			S
			0
			5
			W
			R
	1.6 Testiles		
Syntas	(Check Indent ⊂ Tree View Load Text. Expand Forms Server Query		
Taxonomy @ Depend () () Rules	(∰) Testing [⊕ Data] th_ Document]		
Category Syntax Check		0 X	
Syntax OK			
			• 🖯 🔹
			14:49
Category Syntax Check Build Rulebased Cat	egenzer	I law here it	14-03-2013
Ready		Ln1 NUM	-

Below is a screen shot of the tree view of the same rule.

DEMO.tk2 - SAS Content Categori	ation Studio	Man, 7, Adving inc. Manual King		
Eile Edit View Build Project	Category Concept Testing Document Server Help			
0 🗃 🖬 🙆 🏘 🎱 🌲 🍕	Pa ⊕ ? ⊕ ?			
E 🖗 DEMO	B-SENT		A	
E-Q English	- "business"			
B-@ Top	- "profit"			
Business	"profitable"			
	"break even"			
				-
			S	
			(17)	
			F	
	C Text View	1 1		
	· Tree View	Expand Forms Server Query		
S Taxonomy & Depend ()	4) Rules 4 Testing 3 Data Dh Document			
Category Syntax Check			+ x	
Syntax OK				
			· 🕁 🕪	
			1450	
Category Syntax Check Da Build	Rulebased Categorizer		14-03-2013	3
Ready			Ln1 NUM	

4. On the data tab, set the path of the folder where the files to be tested are located and then click Propagate.

DEMO.tk2 - SAS Content Categorization St	tudio	
File Edit View Build Project Categor	ry Concept Testing Document Server Help	
🗋 🖻 🖬 🙆 🍪 🥔 🌲 🖣 🖻	9 9 9 9	
DEMO ⊕ ⊕ English ⊕ ⊕ Categorizer ⊕ ⊖ Top ⊕ ⊖ Top ⊕ ⊕ Op ⊕ ⊕ Business	ID: <u>Author</u> Created: December 09, 2013 Relevancy Relevancy Category Match Completed Cutoff Bias Bias Ratio C Pending 0 C Test Disabled Description	
		*
	Thesaurus	×
	Comments	×.
		A T
	Related Links	*
	Testing Path Propagate Propagate [C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\categories\Top\Business Propagate IV Tgaining Path Propagate IV IV IV	agate Options dentical Path Create Folders
🖉 Taxonomy 🐼 Dependencies Ready	🚯 Rules 🕼 Testing 🙀 Data 🖾 Document	NUM //

CO V BEMO + test_docs + categories	► Top ► Business ►	255
File Edit View Tools Help		
Organize 🔻 Include in library 👻 Share with	a ▼ Burn New folder	iii • 🗊 🌘
☆ Favorites	Name	Date modified
👪 Abhijit	🐌 Finance	12/9/2013 4:04 PM
🧮 Desktop	📄 B1.txt	12/9/2013 3:58 PM
👪 Downloads	📔 B2-failing-doc.txt	12/9/2013 4:02 PM

DEMO.tk2 - SAS Content Categorization St	tudio	
File Edit View Build Project Categor	ry Concept Testing Document Server Help	
D 🗃 🖬 🙆 🍪 🎱 🌲 🔩 🖕	♥ ♥ ♥ ?	
DEMO Gradient Categorizer Gradient Business	ID: Author Created: December 09, 2013 Modified: Relevancy Relevancy Category Match Completed Bias Ratio C Pending Description	
		×
	Thesaurus	A
	SAS Content Categorization Studio	
		~
	Query Testing paths for subcategories have been set.	
	ОК	~ ~
	Comments	
		*
	Related Links	
		۸ ۳
	Testing Path C\Users\sas\Document\TK240 Projects\Demo\test_docs\categories\Top\Business	agate Options
	Training Path	dentical Path Create Folders
	Propagate	Acade Fonders
Taxonomy 🐼 Dependencies	🚯 Rules 🌗 Testing 🙀 Data 🖻 Document	
Ready		NUM //.

The files in the testing folder will now be visible in the testing tab.

DEMO.tk2 - SAS Content Categorization St	Studio	
File Edit View Build Project Categor	ory Concept Testing Document Server Help	
🛛 🗅 🚔 🖬 🔕 🕸 🥔 🌲 🖣 🐂 🕅	8 8 6 8	
DEMO General English General Categorizer General Top One Development	Test files for this category TEST Crest all files everywhere C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\categories\Top\Business	
Business	Test File Result Relevancy	
	4 Rules ⊕ Testing → Data ⊡ Document	
Ready		M

5. Now, build the category rules we created. To do that, check for the syntax of the rules using Syntax Check before building categories.

DEMO.tk2 - SAS Content Categorization Studio	
File Edit View Build Project Category Concept Testing Document Server Help	
Image: Section of the section of th	
Syntax Check Indent C Tree View Load Text Expand Forms Server Query	
🔯 Taxonomy 🕸 Dependencies 🛛 🚯 Rules 🕼 Testing 📦 Data 🖻 Document	
Category Syntax Check	д X
Syntax OK	
Category Syntax Check	
Ready Ln 1	NUM //

6. Above we created a boolean rule based category called Business, hence we will build the category using Build->Build Rulebased categorizer from the menu bar.

DEMO.tk2 - SAS Cor	ntent Categorization Studio		
File Edit View Bu	ild Project Category Concept T	Testing Document Server Help	
🗋 🗅 🚔 🖬 🕼 🗸	Build Rulebased Categorizer		
DEMO	Build Statistical Categorizer Compile Concepts	ness",(OR,"profit","profitable","break even"))	
	Upload Categorizer Upload Concepts Upload LITI		
	Abort Compiling Concepts		
🗢 Taxonomy 🕸 [Pependencies & Rules	heck Indent C Text View Load Text Expand Forms Server Query ▷ Testing @ Data D Document	
Category Syntax Check			д х
Syntax OK			
Detegory Syntax Ch	eck	[In 1	

Build is successful.

Contraction of the second seco	
DEMO.tk2 - SAS Content Categorization Studio	- • •
File Edit View Build Project Category Concept Testing Document Server Help	
] D 🖆 🖬 Ø � Ø ♣ ♣ ¤ ♥ ♥ ♥ ♥	
DEMO Penglish Categorizer Pop Business (SENT, "business", (OR, "profit", "profitable", "break even")) (SENT, "business", (OR, "profit", "profitable", "break even"))	
Syntax Check Indent C Text View Load Text Expand Forms Server Query	
🔯 Taxonomy 🐼 Dependencies 🛛 🚯 Rules 🐠 Testing 🔯 Data 🖻 Document	
Build Rulebased Categorizer	4 х
Build Successful.	
📳 Category Syntax Check 🔀 Build Rulebased Categorizer	
Ready In 1	NUM //

7. Now to test the files go to the Testing tab and click the Test button.

DEMO.tk2 - SAS Content Categorization St	udio				
File Fdit View Build Project Category Concept Testing Dorument Server Heln					
) (n				
DEMO DEMO Categorizer Demo Categorizer Demo	Test files for this category Test all files everywhere TEST C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\cal	egories\Top\	Business		
E Business	Test File	Result	Relevancy		
Tavonomy @ Dependencies	B1.bt B2-failing-doc.bt	PASS FAIL	0.7778		
) (<u></u>))				
Duild Rulebased Categorizer					4 X
Category Syntax Check 🔒 Build Rulebas	ed Categorizer				
Ready					NUM //

We can double click any of the test documents shown above to see which words get matched or if the test documents did not match. Matched words will get highlighted in red.

Relevancy score shown above is useful to determine best match when the document matches more than one category.

8. Double click the B1.txt to open that file on Document tab and check which words get highlighted.

DEMO.tk2 - SAS Content Categorization S	Studio		83
File Edit View Build Project Catego	ory Concept Testing Document Server Help		
🗋 🗅 🚅 🖬 🙆 🍪 🥔 🌲 🖣 🖣	T		
E DEMO	Test File: C:\Users\sas\Documents\TK240 Projects\Demo\test docs\categ	ories\Top\Business\B1.txt Go Stop	
🛓 🤬 English			
🖻 🆓 Categorizer	4/00000		
E-C Top	success.		
	As per my prediction, his business will break even only in the third year.		
		💷 Rule Matches	
		C. SEAT	
		⊡ OR	
		"profit"	
		"profitable"	
		"break even"	
	PASS TEST © Selected category		
	C All categories		
	All categories and all concepts		
September Cles	Ap Rules Ap Testing Data Document		
Build Rulebased Categorizer			Ψ×
Build Successful.		There are 3 terms matched from 4 terms total.	
		G. Eonward C. Backward Next Match	
		Ser of ward S. Dackward	
関 Category Syntax Check 🔋 Build Ruleba	sed Categorizer		
Ready		Ln 3 NUM	//.

The file *B2-failing-doc.txt* fails because the terms profit and business do not match the rule defined.

DEMO.tk2 - SAS Content Categorization S	tudio			
File Edit View Build Project Catego	ny Concept Testing Document Server Help			
🗋 🖻 🖬 🙆 🕸 🥔 🌲 🖣 🖣	8 9 6 9 8			
E DEMO	Test File: C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\categories\Top	\Business\B2-failing-doc.txt		Go Stop
English	My brother has recently taken a large home loan this year to buy a new home.	······		
⊟ 🎝 Categorizer	He also wanted to purchase a premium car this year with the help of a car loan.			
Business	I know that automobile finance companies will only focus on their profit and will not bother if you	a have any other loans or EMIs going.		
	I thought with both the home and car loan installments it will be a tight financial situation for him For some filendly advice, I suggested to him that he postpones his car buying decision but I am On the contrary, he rudely told me that it is none of my business to look into his personal finance.	afraid that he did not listen to me. es.		
	He will manage his personal finances and will not ask any monetary help from me.			
		Rule Matches		
		"business"		
		⊡ OR		
		"profit"		
		"profitable" "break even"		
	G Salastad estadou			
	C All categories			
	C All categories and all concepts			
🔷 Taxonomy 🛷 Dependencies	🚯 Rules 🕼 Testing 📦 Data 🖻 Document			
Build Rulebased Categorizer				
Build Successful.				
		There is 0 term matched from 4	terms total.	
				Number
🔍 Category Syntax Check 🚯 Build Ruleba	sed Categorizer	(• Forward	C Backward	INEXT MIATCH
Ready			Ln 1	NUM //

9. A good model should match all the relevant test documents (recall) while not matching irrelevant documents (precision). To test the failure of such irrelevant documents we can also check using Testing->Import Failing test Files....

DEMO.tk2 - SAS Content Categorization Str	udio			
File Edit View Build Project Categor	y Concept Te	sting Document Server Help		
File Edit View Build Project Categor Edit View Build Project Categor DEMO D	Y Concept le	string Document Server Help Import Test Files Import Failing Test Files Delete Selected Test Files Save Test Results View Saved Results Export Testing Results To File Populate Testing Paths Restore Populate Results Show Graphical Populate Results Full Test Report Show Last Test Report Graphical Full Test Report Show Last Graphical Full Test Report Show Last Graphical Full Test Report	s\categories\Top\Business\B2-failing-doc.txt e. an. will not bother if you have any other loans or EMIs going. ial situation for him. g decision but I am afraid that he did not listen to me. his personal finances. m me.	Go Stop
	FAIL	TEST C Selected category C All categories C All categories and al	│ Browser View │ View Rule Matches │ concepts	
Taxonomy 🖉 Dependencies	Rules 🐠	Testing 📄 Data 🖻 Document		
Build Rulebased Categorizer				ч х
Build Successful.				
Category Syntax Check Build Rulebase	ed Categorizer			Ln 1 NUM

10. Open the Failing doc to be tested

DEMO.tk2 - SAS Content Categorization	n Studio				
File Edit View Build Project Categ	gory Concept Testing Document Server Help				
🗋 🖻 🖬 🙆 🍪 🎱 🌲 🖣 🖣	1 I I I I I I I I I I I I I I I I I I I				
E- DEMO	Test File: C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\categories\Top\Business\B2-failing-doc.bt	Go Stop 🗮			
English	My brother has recently taken a large home loan this year to buy a new home.				
in Categorizer	He also wanted to purchase a premium car this year with the help of a car loan.				
Business	I know that automobile finance companies will only focus on their profit and will not bother if you have any other loans or EMIs going.				
	I thought with both the home and car loan installments it will be a tight financial situation for him. For some friendly advice, I suggested to him that he postpones his car buying decision but I am atraid that he did not listen to me				
	Look in: 👔 Business				
	Name Date modified				
	Recent Places Finance 12/9/2013 4:04 PM				
	B1.bt 12/9/2013 3:58 PM				
	ta status.xml 2/13/2013 10:42 Alv				
	Desktop				
	Libraries				
	Computer is D76048				
	Notwork Elements D2 (stimutes b)				
🖉 Taxonomy 🖉 Dependencies	Files of type: Test Files (".txt, ".ttf, ".html, ".html, ".sgml, ".xml) Lancel				
Build Rulebased Categorizer		д X			
Build Successful.					
Category Syntax Check Build Rulebased Categorizer					
Ready		1 NUM			

The imported file will be visible in the Testing tab.

DEMO tk2 - SAS Content Optoporization St	tudio	
File Edit View Build Broject Categori	uuluo	
	יזי סטוובקא ובעווק בסכטוובוג בויער וובף או או או או או	
	 Test files for this category 	
Categorizer	C Test all files everywhere IEST	
⊡⊕ Top	C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\categories\Top\Business	
Business	Test file Result Relevancy	
	fail\82-failing-doc.txt	
	B2-failing-doc.txt	
	prot	
🗇 Taxonomy 🐼 Dependencies	🚯 Rules 🕼 Testing 🔯 Data 🖻 Document	
Build Rulebased Categorizer		4 X
Build Successful.		
Category Syntax Check 🚯 Build Rulebase	ved Categorizer	
Ready		1

In the Test the imported failing test file should FAIL, that proves our rule is precise (filtering irrelevant documents) and working correctly.

DEMO.tk2 - SAS Content Categorization St	udio					
File Edit View Build Project Category Concept Testing Document Server Help						
🗋 🗅 🚅 🖬 🙆 🍪 🥔 🌲 🖣 🖡 👌	9 49 (9 9					
DEMO English Categorizer Or Top Business	 ☞ Test files for this category 	egories\Top\ Result PASS FAIL FAIL	Business Relevancy 0.7778			
Taxonomy & Dependencies	🔐 Rules 🖓 Testing 🔤 Data 🖻 Document					
Build Rulebased Categorizer						₽ X
Build Successful.	ed Colonarian					
	a Categorizer					
Ready						INUM I

11. Double clicking the Failing document opens it in the Document Tab.

		- tota taka theory takata the				
DEMO.tk2 - SAS Content Categorization S	tudio					
File Edit View Build Project Catego	ry Concept Testing Document Server Help					
📄 🗅 🚅 🖬 \mid 🔕 🗇 🥔 ≱ 🖣 🖣 '	T & & T					
E-@ DEMO	Test File: C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\categor	ies\Top\Business\fail\B2-failing-doc.txt	Go Stop 🗮			
English	My brother has recently taken a large home loan this year to buy a new home.					
⊟	Eragi Categorizer He also wanted to purchase a premium car mis year with the help of a car toan.					
Business	I know that automobile finance companies will only focus on their profit and will not bol	her if you have any other loans or EMIs going.				
	I thought with both the home and car loan installments it will be a tight financial situation. For some friendly advice, I suggested to him that he postpones his car buying decision Ω n the contrary, he rudely told me that it is none of my business to look into his person	n for him. but I am afraid that he did not listen to me. al finances.				
	He will manage his personal finances and will not ask any monetary help from me.					
		Rule Matches				
		SENT				
		"business"				
		⊡-OR				
		"profit" "profitable"				
		"break even"				
	FATI FATI • Selected category					
	C All categories					
A T A Dependencies	All categories and all concepts					
Taxonomy so Dependencies	<u>ar rules</u> <u>ar resurg</u> <u>Data</u> <u>Document</u>					
Build Rulebased Categorizer						
Build Successful.						
		There is 0 term matched from 4 terms total.				
		Forward C Backward	Next Match			
-						
Category Syntax Check 🔀 Build Rulebas	ed Categorizer					
Ready			Ln 1 NUM			

Part 3: Creating LITI Definition Concepts

1. Right mouse click on English and then click Enable Concepts to create Concept Taxonomy



DEMO.tk2 - SAS Content Categorization Studio	
File Edit View Build Project Category Concept Testing Document Server Help	
D 🖆 🖬 🕼 🕸 🚳 🔺 🖣 🗣 🗇 🗇 🗇 😤	
Concepts Top	
Transana & Dependencies & Ab Definition & Testing & Data Db Document	
Puller Vulebaseo, Sergorizer	4 ×
Build Successful.	
🔁 Category Syntax Check 🚯 Build Rulebased Categorizer	
Ready	NUM //

2. To add new Concept do Right Click on Top and select Add Concept

The Demonstrate Association Studio	
The tait view suila Project Category Concept Testing Document server Help	
Add Concept	
Paste	
Paste Single Node	
Paste Symbolic Link	
Expand Fully	
🗇 Taxonomy 🐼 Dependencies 🛛 🚯 Definition 🌗 Testing 🗊 Data 📴 Document	
Julid Rulebased Categorizer	4 X
Build Successful	
👷 Category Syntax Check 🙀 Build Rulebased Categorizer	
Ready	NUM //

3. Create a concept definition named liti_regex_email. In the below screenshot the LITI concept uses Regular Expressions to capture the email address ending with .com. To create the LITI concept, select the LITI radio button on the Definition tab.

DEMO.tk2 - SAS Content Categorization Studio	
File Edit View Build Project Category Concept Testing Document Server Help	
DEMO ⊕ © English ⊕ ⊘ Categorizer ⊕ ⊘ Toncepts ⊕ ⊗ Itti_regex_email	
🔯 Taxonomy 🐼 Dependencies 🔹 🚯 Definition 🕼 Testing 🙀 Data 🖻 Document	
Build Rulebased Categorizer	4 х
Build Successful.	
🔁 Category Syntax Check 🚯 Build Rulebased Categorizer	
Ready	Ln 1 NUM

Write the definition

REGEX: $[\w\.\-] + @ [\w\.\-] + .com$

on the Definition tab and then check for the syntax by clicking Syntax Check button.

DEMO.tk2 - SAS Content Categorization Str	ıdio	
File Edit View Build Project Category	y Concept Testing Document Server Help	
🗋 🖻 🖬 🔗 🍪 👙 🖣 🖣	0 0 0 0	
OEMO Genglish A Categorizer Gorcepts Gorcepts	REGEX: [\w\.\-]+8[\w\.\-]+.com	
	Syntax Check Grammar C LTT Load Text	
🖉 Taxonomy 🐼 Dependencies	🚯 Definition 🕼 Testing 📦 Data 🖬 Document	
Concept Syntax Check		4 X
Syntax OK		
関 Category Syntax Check 🔒 🚯 Build Rulebase	d Categorizer 🛛 🤯 Concept Syntax Check	
Ready	ln :	1 NUM //

4. On the Data tab set the path of the folder where the files to be tested are located and then click Propagate.

DEMO.tk2 - SAS Content Categorization	Studio				
File Edit View Build Project Categ	jory Concept Testing Docume � ゆ み (ター) ター	nt Server Help			
English	ID:	Author	Created: December 09, 20	013	
E Sa Concepts			Modified:		
i⊡- 🏈 Top — 🎸 liti_regex_email	Relevancy Priority	 Completed 	C Case Sensitive Matchi	ng	
	10	C Pending C Test Disabled	 Case Insensitive Match Use Project Default 	hing	
	Description				
	Thesaurus				
					*
	Query	S Content Categorization Studio	X		
	Comments	A Testing naths for subconcents ha	ve heen set		÷
			ve been seu		4 2
	Related Links				
	Testing Dath		ОК		* *
	C:\Users\sas\Documents\TK2	240 Projects\Demo\test_docs\concepts		Propag	ate Propagate Options
					Create Folders
Truncias	A Definition A Testing	Dete Die Document			
appendencies		Data Document			
Svntax OK					
🖏 Category Syntax Check 🔀 Build Ruleba	ased Categorizer 🛛 🤯 Concept Synta	x Check			NUM
,				I	,
😋 🔾 🗢 📕 « TK240 F	Projects ► DEMO ► to	est_docs ► concepts	→ 4	• Search concepts	
File Edit View Too	ols Help				
Organize 🔻 Include	in library 🔻 Share	with 🔻 Burn New	folder		
🚖 Favorites		Name			Date modified
퉬 Abhijit		📋 239036newsML.t	xt		1/22/2010 11:55 AM
📃 Desktop		239040newsML.t	xt		1/22/2010 11:55 AM
鷆 Downloads		📋 239153newsML.t	xt		1/22/2010 11:55 AM
🗐 Recent Places		239156newsML.t	xt		1/22/2010 11:55 AM
🌗 Testfil		239157_sports.xr	nl		1/22/2010 11:55 AM
		89856467 fact.tx	t		1/22/2010 11:55 AM
詞 Libraries		test-doc1.txt			1/22/2010 11:55 AM
Documents		test-doc2.txt			1/22/2010 11:55 AM
J Music		test-doc3 tyt			1/22/2010 11:55 AM
					1, 22) 2010 11:00 AIVI

Files in the Testing Folder will now be visible in the Testing Tab.

DEMO.tk2 - SAS Content Categorization St	tudio	• X
File Edit View Build Project Categor	ry Concept Testing Document Server Help	
🗋 🖻 🖨 🔗 🗞 🁙 🖣 🖣	\$ \$ \$ \$	
OEMO English Categorizer Concepts Co	Cast files for this concept Test files everywhere Test files everywhere Test file Concordance for Selected Concept Concordance for All Concepts Concordance for All Concepts Test file Result Above Relevancy Cutoff test-doc3.tct test-doc3.tct 239155newsML.tct 239155newsML.tct 239030newsML.tct	
🖉 Taxonomy 🖉 Dependencies	🚯 Definition 🕼 Testing 🔯 Data 🖻 Document	
Concept Syntax Check		ąχ
Syntax OK		
Category Syntax Check 🚯 Build Rulebase	sed Categorizer 📴 Concept Syntax Check	
Ready	NUM	

5. Compile the concepts by clicking Build->Compile Concepts on the menu bar

DEMO.tk2 - SAS Con	ntent Categorization Stud	dio						
File Edit View Bui	ild Project Category	Concept Testin	g Document Se	erver Help				
🗅 🚔 🖬 🙆 🗸	Build Rulebased Categ	gorizer						
E DEMO	Build Statistical Catego	orizer			No Concordan	re		
🗄 ᡇ English	Compile Concepts	rtn	is concept	TEST	Concordance f	or Selected Concept		
🕀 🎝 Cate		ev	erywhere		Concordance f	or All Concepts		
i i i i i i i i i i i i i i i i i i i	Upload Categorizer	icu	ments\TK240 Proje	cts\Demo\test_docs	\concepts	1	1	
	Upload Concepts	_			Result	Above Relevancy Cutoff		
	Upload LITI							
	Abort Compiling Conc	cepts						
		89850407_ract.txt						
		239157_sports.xml 239156newsML txt						
		239153newsML.txt						
		239040newsML.txt						
	2	239036newsML.txt						
		A Decision A		ا ما م				
S Taxonomy	j	As Deliving of	Plesting Be Dat					
Concept Syntax Check								д X
Syntax OK								
Di cata angla cata a chu		Cotominan 🗖						
Category Syntax Cha	eck j GM Build Rulebased	Categorizer 🛛 🐺 Ci	oncept Syntax Check	9				
								NUM //

LITI concept build is successful.

DEMO.tk2 - SAS Content Categorization St	tudio	- • ×
File Edit View Build Project Catego	ry Concept Testing Document Server Help	
🗋 🗅 😅 🖬 🙆 🕸 🎱 🌲 🖣 📭 1	\$ \$ \$ \$	
Categorizer Concepts Concepts	REGEX: [\w\.\-]+8 [\w\.\-]+.com	
Taxonomy 🐼 Dependencies	& Definition Data Dh Document	
Compile Concents		
Built second intermediate configuration file (step 5 Built Internal concept matcher (step 5 of 9). Finished building concepts (step 7 of 9). Contextual extraction support detected; preparing Building LITI concepts, please wat LITI concepts build successful Finished removing temporary files (step 8 of 9). Finished removing temporary files (step 8 of 9).	5 of 9). I to build LITI concepts.	
関 Category Syntax Check 🗍 🔀 Build Rulebas	ed Categorizer] 😓 Concept Syntax Check _ 🔯 Compile Concepts	
Ready	ln1	NUM //

6. Now test the files for this concept by clicking the Test button on the Testing tab. With Concordance selected, we can control the number of characters, words or sentences displayed before and after each match.



7. We can double click any of the test documents shown above to see which strings get matched. Matched strings will get highlighted in red as shown below.

Double click the 239157_sports.xml to open the file on Document tab and check which words get highlighted.

DEMO tk2 - SAS Content Categorization S	tudio	
File Edit View Build Project Catego	ny Concent Tecting Document Server Heln	
	δ) (δ) (δ) (δ) (δ)	
	Test File: C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\concepts\239157_sports.xml	<u>60</u> Stop E
	//SubjectCode>	A
i An Categorizer	<pre></pre>	
	<newscomponent equivalentslist="no" xml:lang="en"></newscomponent>	
⊡-≪⊘ lop	<bole formalname="Main"></bole>	
iti_regex_email	(NewsLines) A Head inst Need For Blinded Payment In Bandomized Trials is Benomized.	
	In Recent Bills/HeadLine>	
	<byline>Gregory Twachtman (<u>a.twachtman@elsevier.com</u>)</byline>	
	<dateline>August 18, 2008</dateline>	
	<countine 2008<="" copyrightline="" exposite="" inc.="" reports="" tdc="" to=""></countine>	
	<contentitem></contentitem>	
	<pre><media1ype+romainame=1ext></media1ype+romainame=1ext> </pre>	-
	<datacontent></datacontent>	=
	<pre></pre> <pre><</pre>	
	"http://www.w3.org/TB/xhtml1/DTD/xhtml1-strict.dtd">	
	<html xmlns="http://www.w3.org/1999/xhtml"></html>	
	<pre>chead> chead></pre>	
	 Aitie> 	
	(body)	
	Recent legislation is recognizing the need to blind patients	-
	Selected concert Concordance Review	
	2 Matches TEST Collection of the Concept Conce	
	All categories and all concents	
V Taxonomy W/ Dependencies	<u>A v Dennidon</u> <u>Av Tesung</u> <u>an Data</u> <u>Document</u>	
Compile Concepts		÷ ×
Built second intermediate configuration file (step s	5 of 9).	
Built internal concept matcher (step 6 of 9).		
Contextual extraction support detected: preparing	to build LTI concepts	
Building LITI concepts, please wait		_
LITI concepts build successful.		=
Finished removing temporary files (step 8 of 9).		
Finished. Now freeing resources (step 9 of 9).	and Onteresting and Bill Constant Contast Character and a second s	•
Category Syntax Check GB Build Rulebas	ed Lategorizer 🔯 Concept Syntax Check 🛛 🗳 Compile Concepts	
Ready		Ln 112 NUM

A good model should match all the relevant test documents (recall) while not matching irrelevant test documents (precision).

8. To test the failure of such irrelevant documents which do not satisfy the LITI definition of the concept, we can check by clicking Testing->Import Failing test Files... on the menu bar.

DEMO.tk2 - SAS Content Categorization Studio			- • •				
File Edit View Build Project Category Concept	Testing Document Server Help						
🗋 🗅 🚅 🖬 🔕 🕸 🎱 🔺 🖣 🖣 👻 🏵 🗇	Import Test Files						
DEMO Test File:	Import Failing Test Files	s\concepts\239157 sports.xml	Go Stop 🗮				
English	Delete Selected Test File						
🗄 🆓 Categorizer	Saus Task Basults						
E S Concepts	Save Lest Results						
	view saved Results						
In Recent	Export Testing Results To File	ials Is Recognized					
In neceni	Populate Testing Paths	r.com)					
	Bestore Populate Besults						
	Show Graphical Populate Results	ightLine>					
		-					
	Full Test Report						
	Show Last Test Report		E				
	Graphical Full Test Report		_				
br "http://www	Show Last Graphical Full Test Report						
<hr/> (htmi xmms <head) <title> <td>"http://www.w3.org/1999/xhtml"></td><td></td><td></td></title></head) 	"http://www.w3.org/1999/xhtml">						
<pre></pre>							
Recent	egislation is recognizing the need to blind patients		*				
2 Match	es TEST • Selected concept	🗖 Concordance 🔲 Browser View					
	C All concepts C All estension and all as	View Rule Matches					
Con T Concentencies	→ Air categories and air co	incepts					
	don w resung Data Document						
Compile Concepts			. т х				
Built second intermediate configuration file (step 5 of 9). Built internal concept matcher (step 6 of 9).			^				
Finished building concepts (step 7 of 9).							
Contextual extraction support detected: preparing to build LITI co Building LITI concepts, please wait	incepts.						
LITI concepts build successful.							
Finished, Now freeing resources (step 9 of 9).			*				
🕘 Category Syntax Check 📮 Build Rulebased Categorize	🖞 😺 Concept Syntax Check 📑 Compile Concepts						
			Ln 112 NUM				

Open the failing document *fail_email_test.txt* to be tested as shown below

DEMO.tk2 - SAS Content Categorization St	udio				
File Edit View Build Project Categor	y Concept Testing [ocument Server Help			
🗋 🗅 🚅 🖬 🕼 🕸 🎱 🗍 🗛 📭 t	9999				
E-♀ DEMO E-₩ English E-♪ Categorizer E-♀ Concepts	Test files for this co Test all files everyw C:\Users\sas\Documen	ncept C No C here TEST C Conc C Conc ts\TK240 Projects\Demo\test docs\concer	Concordance cordance for Selected Conc cordance for All Concepts ots	cept	
- 🖉 Top	Test File	F	Result Above Releva	ncy Cutoff	
iti_regex_email	239036 239040 239153 Look ir				
	239156 239157 299564 test-dc Fecent Places test-dc Desktop Libraries	Name 239036newsML.bt 239040newsML.bt 239153newsML.bt 239157_sports.xml 98956467_fact.bt fail_email_test.bt test-doc1.bt test-doc2.bt test-doc3.bt	· • • • •	Date modified 1/22/2010 11:55 AN 1/22/2010 11:55 AN	
🗢 Taxonomy 🐼 Dependencies	Network	File name: fail_email_test.txt Files of type: Test Files (".txt, ".rtf,".h	tml, [*] .htm, [*] .sgml, [*] .xml)	Dpen Cancel	
Compile Concepts		,			4 X
Built second intermediate configuration file (step 5 Built internal concept matcher (step 6 of 9). Finished building concepts (step 7 of 9). Contextual extraction support detected, preparing Building LIT concepts, please wat LITI concepts build successful. Finished removing temporary files (step 8 of 9). Finished. Now freeing resources (step 9 of 9).	of 9). to build LITI concepts.				A H V
Category Syntax Check 💀 Build Rulebas	ed Categorizer 🛛 😓 Concep	ot Syntax Check 🛛 💱 Compile Concepts			
Ready					NUM //.

The imported file will be visible in the Testing tab.

📓 DEMO.tk2 - SAS Content Categorization St	udio	
File Edit View Build Project Categor	y Concept Testing Document Server Help	
🗋 🗅 🚅 🖬 🕼 🍪 🎱 🔺 🖣 📭 t	\$ \$ \$ \$	
DEMO Calegorizer Concepts Concept	Image: Construction of the second	
Taxonomy 🐼 Dependencies	🚯 Definition 🕼 Testing 📦 Data 🖻 Document	
Compile Concepts		4 X
Built second intermediate configuration file (step 5 f Built internal concept matcher (step 5 of 9). Finished building concepts (step 7 of 9). Contextual extraction support detected, preparing Building LIT concepts, please wath LITI concepts build successful. Finished removing temporary files (step 8 of 9). Finished. Now freeing resources (step 9 of 9).	of 9). to build LITI concepts.	E
Ready	ea Categorizer 🥪 Concept Syntax Check 🔮 Compile Concepts	
ricouy		

In the test, the imported file fail_email_test.txt should not match any strings. You can verify this by noticing that the value for the Result column is 0. This shows that our concept definition is precise (filtering irrelevant documents) and working as expected.

DEMO.tk2 - SAS Content Categorization St	udio		and the table france to also	
File Edit View Build Project Categor	v Concept Testing Document Server Help			
🗍 🗅 🚔 🔛 🔕 🍪 🥝 🔺 🗛 📭 🕈	\$ \$ \$ \$ \$			
DEMO © English © Categorizer © Concepts © Top © Itt_regex_email	C Test files for this concept C Test files everywhere CAUsers\iss\Document\TK240 Project\Demo\test_docs\co Test file 239157_sports.xml test-doc1.bt fail_email_testbt fail_email_testbt fail_email_testbt fail_fail_entestbt 239156newsML.bt 239040newsML.bt	No Concordance Concordance oncepts 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	nce for Selected Concept for All Concepts 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
🖉 Taxonomy 🖉 Dependencies	🚯 Definition 🕼 Testing 🔯 Data 🖻 Document			
Compile Concepts				д х
Built second intermediate configuration file (step 5 Built internal concept matcher (step 6 of 9). Finished building concepts (step 7 of 9). Contextual extraction support detected; preparing Building LIT concepts, please wat LITI concepts build successful. Finished removing temporary files (step 8 of 9). Finished. Now freeing resources (step 9 of 9).	of 9). to build LITI concepts.			
Ready	ed Categorizer J 🥹 Concept Syntax Check 🛛 🕎 Compile Concepts	<u>.</u>		NUM
neauy				

No strings get highlighted for the file fail_email_test.txt as it does not satisfy our REGEX LITI Definition

DEMO.tk2 - SAS Content Categorization S	tudio	
File Edit View Build Project Catego	ry Concept Testing Document Server Help	
🗋 🗅 🚅 🖬 🙆 🕸 🍪 🌲 🖣 🐂	T	
E-Q DEMO	Test File: C:\Users\sas\Documents\TK240 Projects\Demo\test_docs\concepts\fail\fail_email_test.txt	Go Stop 🗮
⊡	He also clarifiedd to Professor Hall by writing to k.hall@sample.edu.	
🗈 🖓 Categorizer	Other failing email addresses are as follows -	
Concepts	xyz@news.biz	
iti regex email	John.K@gethelp.org	
	Matches TEST Selected concept Concordance Browser View	
	C All concepts View Rule Matches	
Tauranamu @ Dependencies	All Definition All Testing All Data and Decument	
Taxonomy to bependencies		
Compile Concepts		.
Built second intermediate configuration file (step 5 Built internal concept matcher (step 6 of 9).	o 01 3).	^
Finished building concepts (step 7 of 9).	s to by Hill 1991 occurrents	
Building LITI concepts, please wait	puo auna Lin Concepts.	
LITI concepts build successful.		=
Finished. Now freeing resources (step 9 of 9).		-
🕘 Category Syntax Check 関 Build Rulebas	ed Categorizer 🛛 🧓 Concept Syntax Check 📑 Compile Concepts	
Ready		Ln 1 NUM