ASK THE EXPERT

Getting Started with SAS® Visual Text Analytics

George Fernandez
George.Fernandez@sas.com

Copyright © SAS Institute Inc. All rights reserved.

George Fernandez

Senior Analytical Training Consultant, SAS

George has more than 30 years of experience in teaching advanced analytics courses. He leads training in predictive modeling, forecasting, text analytics, model management and survey analysis.

Copyright © SAS Institute Inc. All rights reserved.

AGENDA

- *Visual Text Analytics – A gentle introduction*
- *Visual Text Analytics – Steps*
- *Visual Text Analytics – Features*

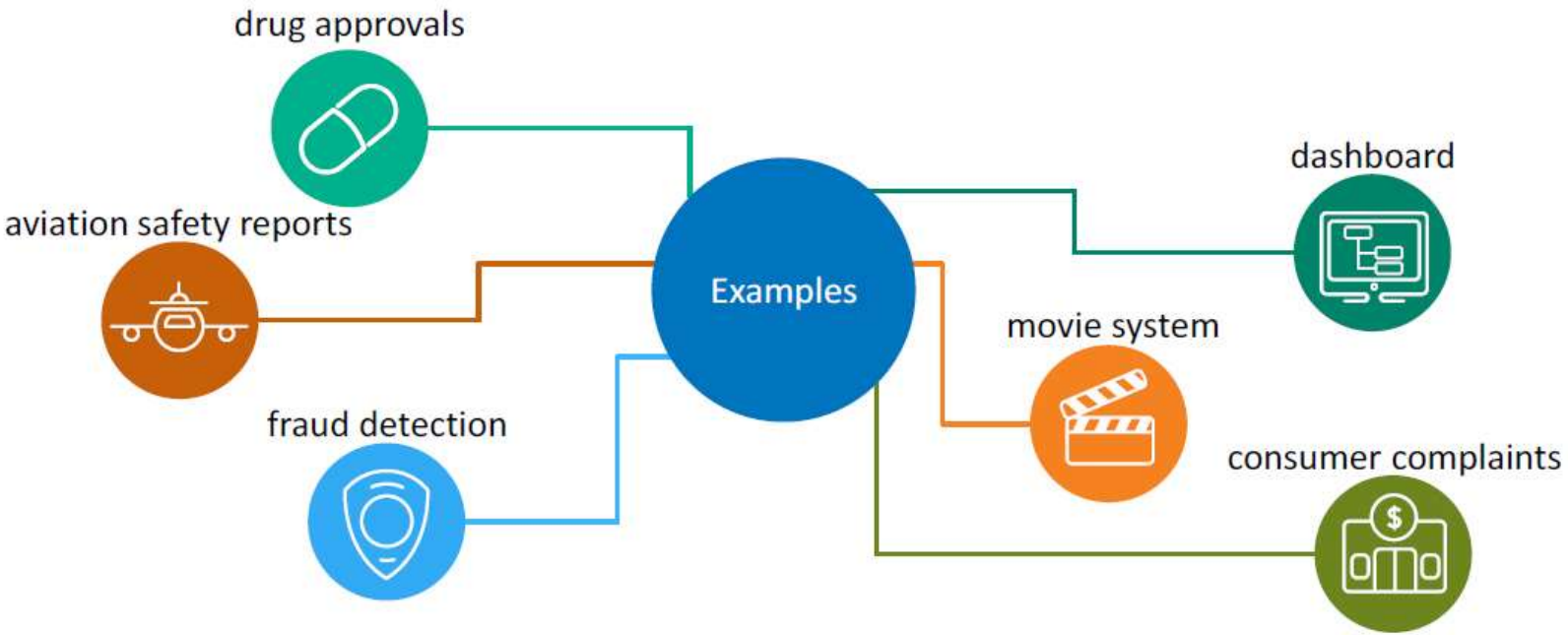
Copyright © SAS Institute Inc. All

SAS Visual Text Analytics

SAS Visual Text Analytics combines the power of natural language processing, machine learning, and linguistic rules in an end-to-end analytics framework with the goal of creating and refining models for extracting insights from textual data.

Copyright © SAS Institute Inc. All rights reserved.

Text Analytics: Examples



Copyright © SAS Institute Inc. All rights reserved.

Why Is SAS Visual Text Analytics Important?

SAS Visual Text Analytics automatically converts unstructured data into meaningful insights. Considering the staggering amount of unstructured data that's generated every day, automation is critical to fully analyze text and speech data efficiently.

- Sepsis Indicators
- Painkiller Addict
- Environmental Influences

- Is product meeting quality expectations?

- Terrorism
- Cyber Bullying
- At-Risk Persons

- What amenities influence bookings?

25

sas

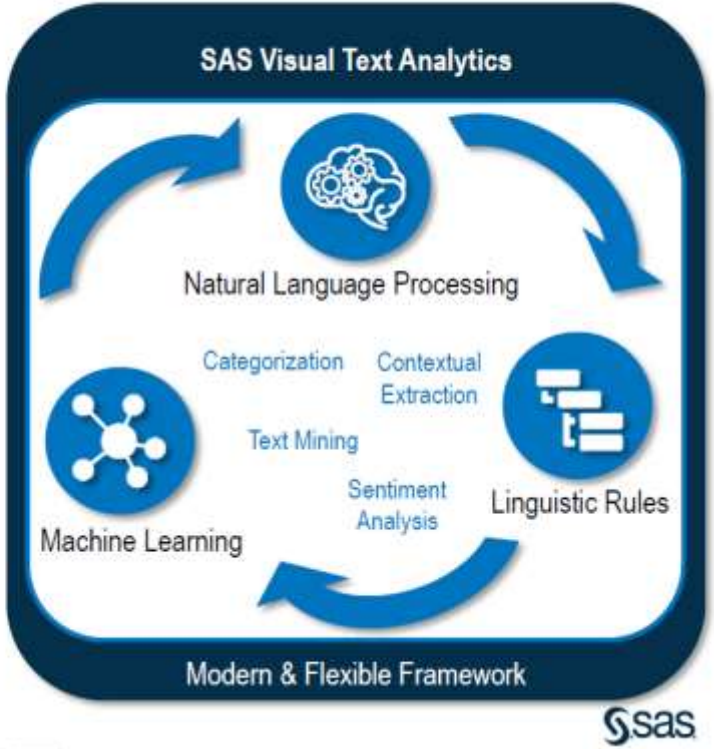
Copyright © SAS Institute Inc. All rights reserved.

What Is SAS Visual Text Analytics?

SAS Visual Text Analytics is the SAS offering designed to effectively extract insights from unstructured data in large scale.

Offered on the SAS Viya architecture, Visual Text Analytics combines the power of

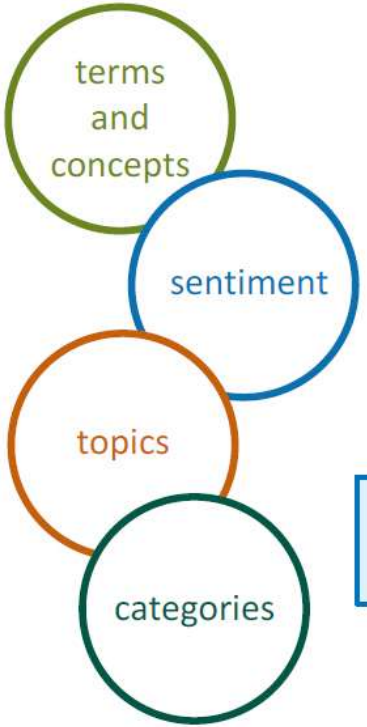
- natural language processing (NLP)
- machine learning (ML)
- linguistic rules (algorithmically derived or human supplied).



22

Copyright © SAS Institute Inc. All rights reserved.

SAS Visual Text Analytics: Big Picture



SAS Visual Text Analytics

SAS Viya

Model Studio

Cloud Analytic Services (CAS)

Copyright © SAS Institute Inc. All rights reserved.

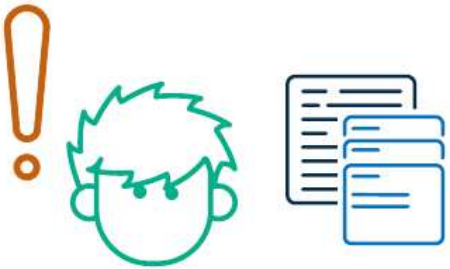
Global Language Support

33 Languages

Arabic	Greek	Portuguese
Chinese	Hebrew	Romanian
Croatian	Hindi	Russian
Czech	Hungarian	Slovak
Danish	Indonesian	Slovene
Dutch	Italian	Spanish
English	Japanese	Swedish
Farsi	Kazakh	Tagalog
Finnish	Korean	Thai
French	Norwegian	Turkish
German	Polish	Vietnamese

Copyright © SAS Institute Inc. All rights reserved.

Why Use SAS Visual Text Analytics?



eliminates manual reviews

removes manual bias errors

regulates consistent classification

removes requirement of training corpus

Copyright © SAS Institute Inc. All rights reserved.

Unstructured Text to Structured Data

unstructured

text data

- information overload
- cognitive bias
- sampling bias
- automatic knowledge or intelligence discovery
- multiple languages and cultures
- limited information lifespan

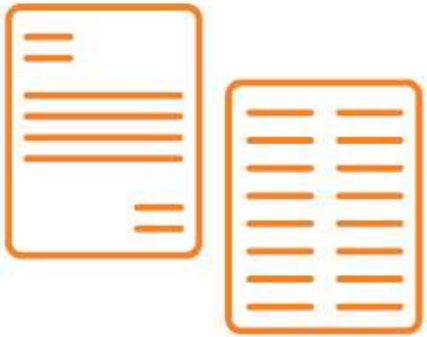
- document collections are massive
- text parsing is computationally intensive
- decisions about strategic versus tactical time frames
- real-time knowledge capture; real-time deployment problematic

structured

Copyright © SAS Institute Inc. All rights reserved.

Text Analytics Projects Requirements

well-defined goal



document collection

metrics to identify success

Copyright © SAS Institute Inc. All rights reserved.

Text Analytics

data = documents

Text Analytics Project

natural language processing (NLP)

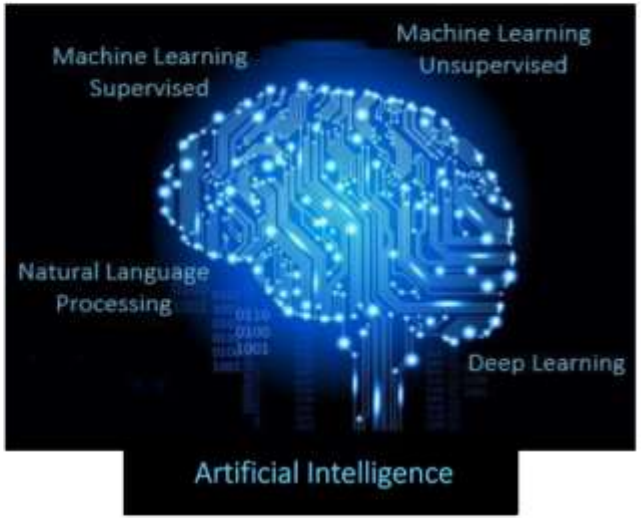
collection of dictionaries

Copyright © SAS Institute Inc. All rights reserved.

What Is Natural Language Processing (NLP)

Branch of artificial intelligence that helps computers

- understand
- interpret, and
- manipulate human language.



NLP helps **resolve ambiguity** in language and **adds useful numeric structure** to the data for many downstream applications, such as speech recognition or text analytics.

Copyright © SAS Institute Inc. All rights reserved.

Powerful Algorithms and NLP

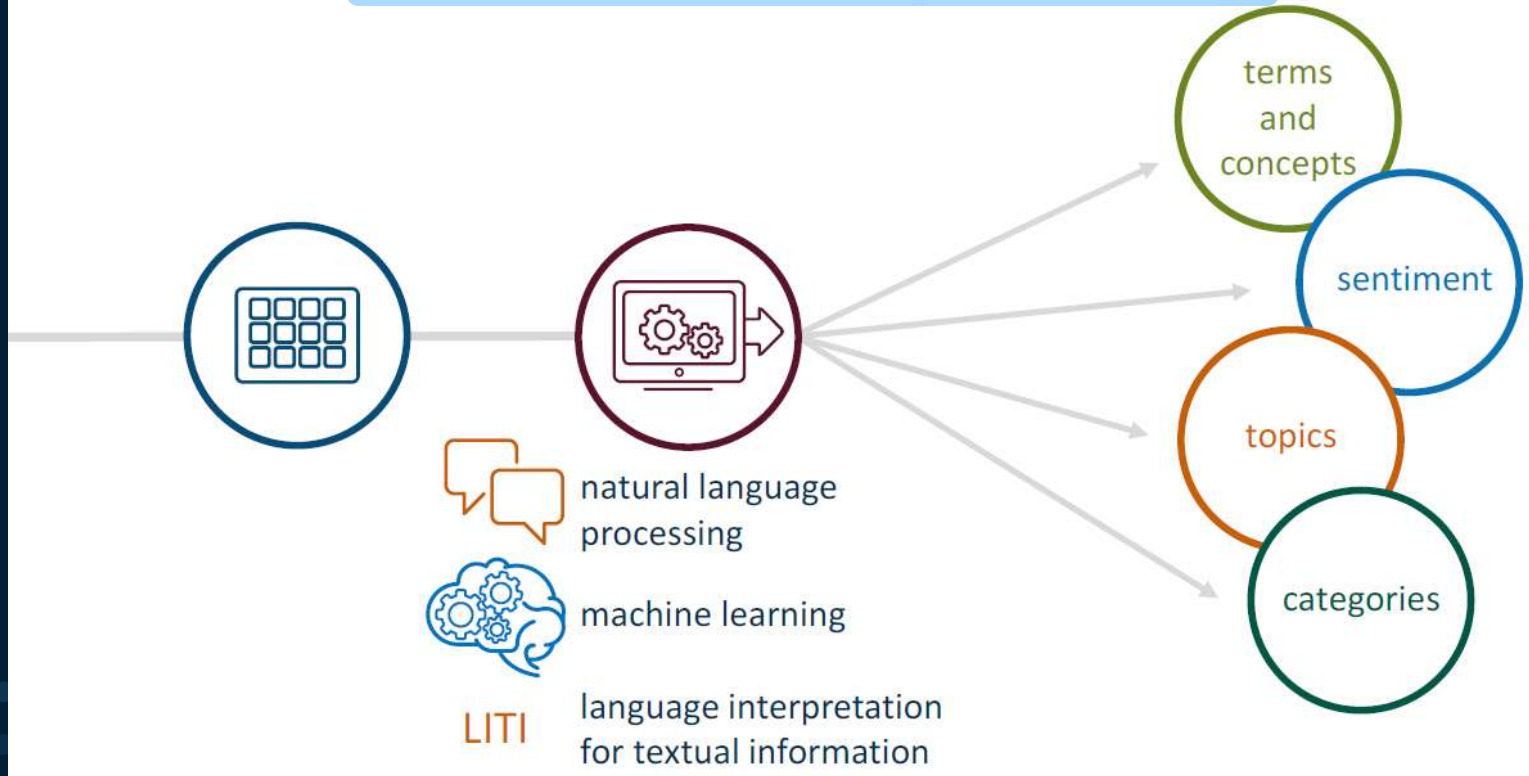
Intelligent algorithms and natural language processing techniques automatically extract relationships and patterns.

Including:

- stemming
- part-of-speech tagging
- compound decomposition
- segmentation
- named entity recognition
- semantic parsing
- tokenization and sentence breaking

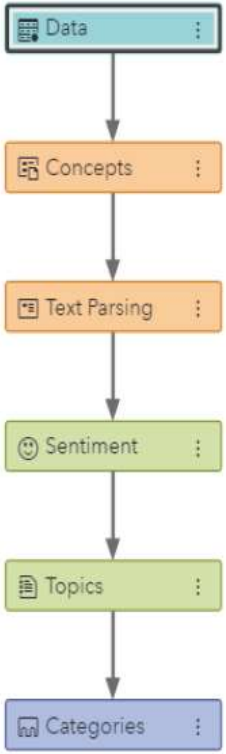
Copyright © SAS Institute Inc. All rights reserved.

SAS Visual Text Analytics: Process



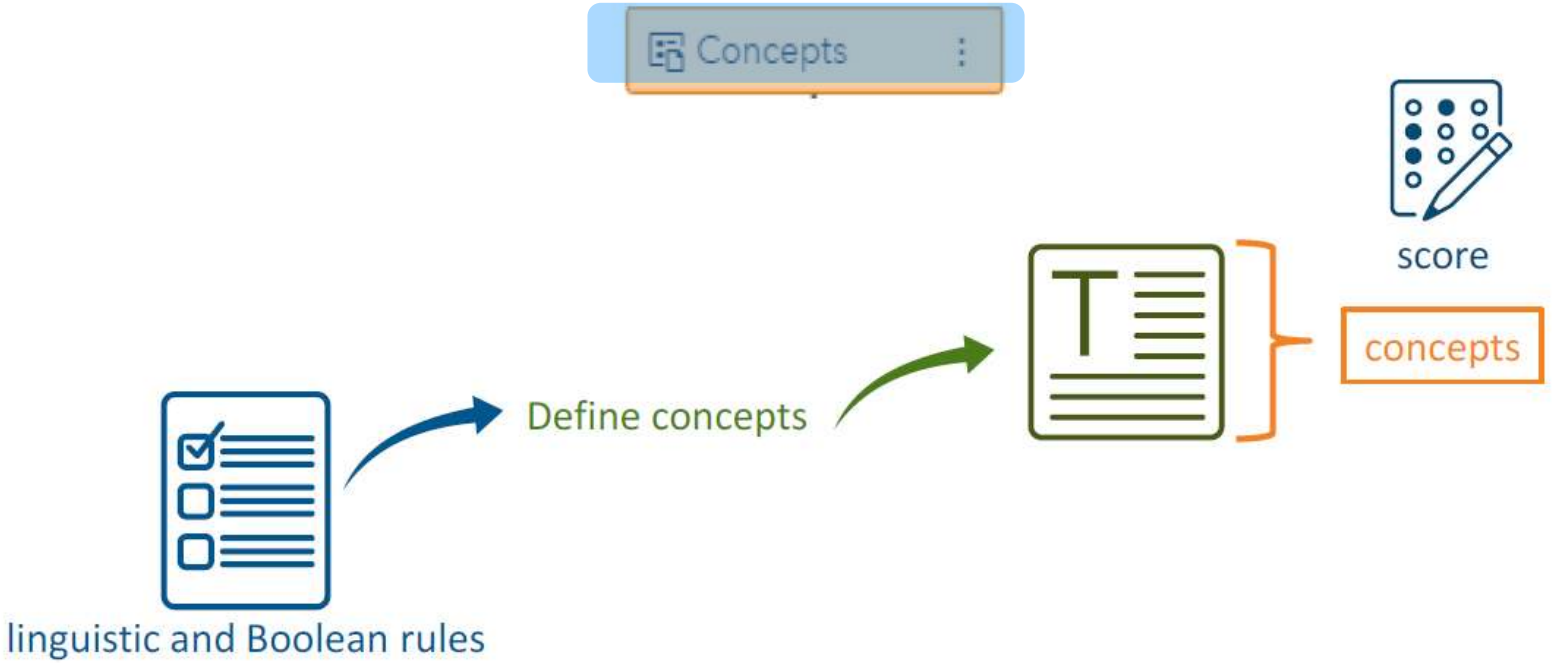
Copyright © SAS Institute Inc. All rights reserved.

SAS Visual Text Analytics: Default Pipeline



Copyright © SAS Institute Inc. All rights reserved.

Text Analytics Components



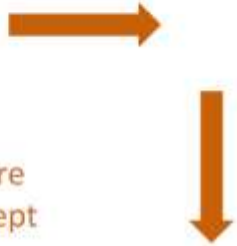
Copyright © SAS Institute Inc. All rights reserved.

Concept Extraction

Input text data
(Twitter data, reports, email, news, or forum messages)

Parse defined concepts (linguistic and Boolean rules)

Output results (scored using score code; satisfy concept definitions)



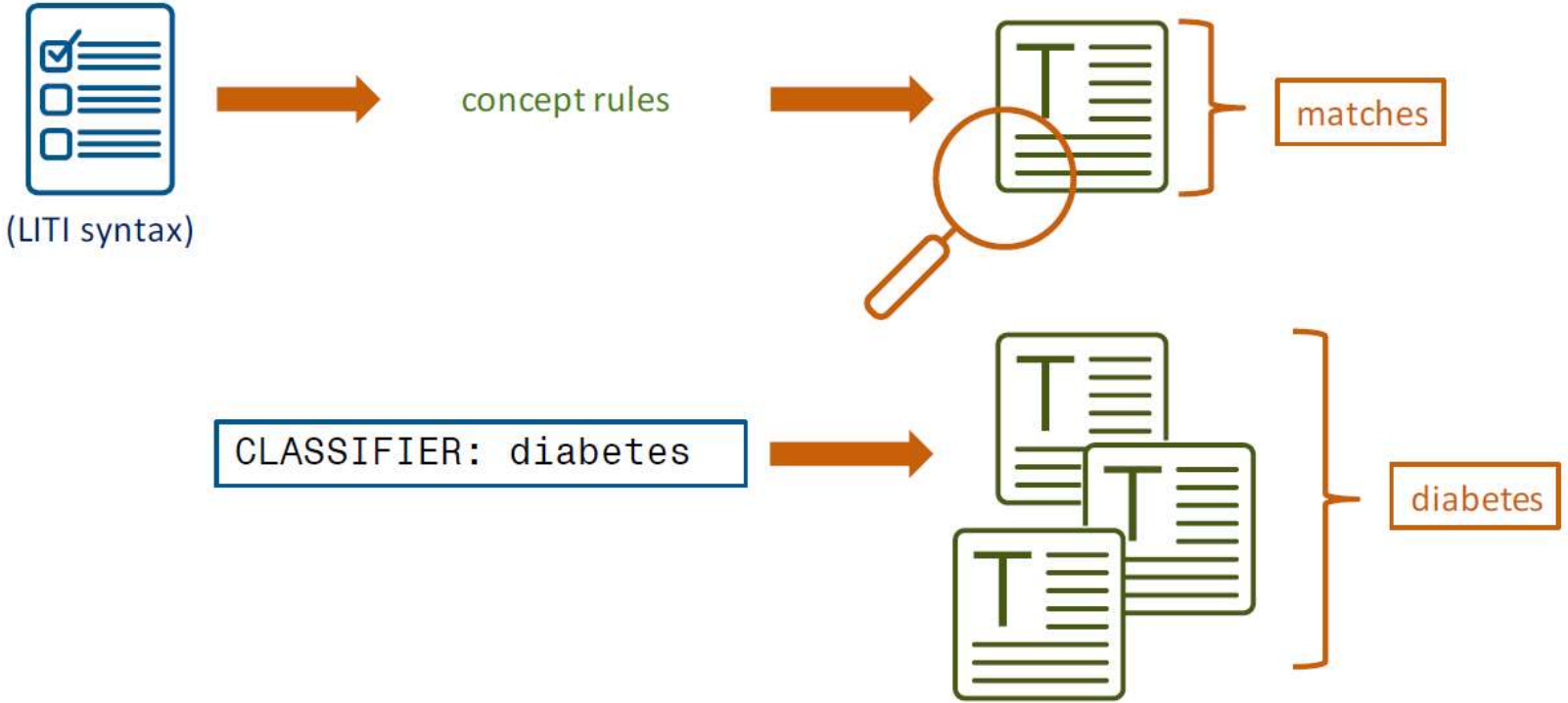
Research and analysis

Concepts

- Measures – 50 mg…
- Persons – John…
- Dates – Monday…
- Organizations – SAS…

Copyright © SAS Institute Inc. All rights reserved.

Introduction to Concept Rules



Copyright © SAS Institute Inc. All rights reserved.

Text Analytics Components

```
1  REGEX:[0-9]+[\.][0-9]+\s?mg\.?
2  REGEX:[\d]+\s?mg\.?
```

I have been on **10mg** prexifan for 6 months along with **2000mg** noderall **1mg** gemulex and fortifex when needed.

Copyright © SAS Institute Inc. All rights reserved.

Concept Rules

Extracting Concepts:

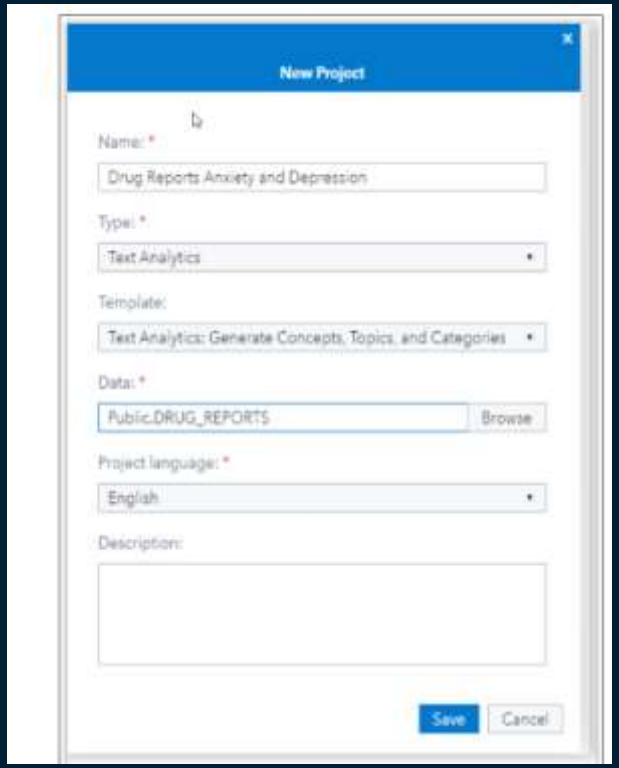
- CLASSIFIER
- CONCEPT
- C_CONCEPT
- CONCEPT_RULE
- NO_BREAK
- REGEX
- REMOVE_ITEM

Extracting Facts:

- PREDICATE_RULE
- SEQUENCE

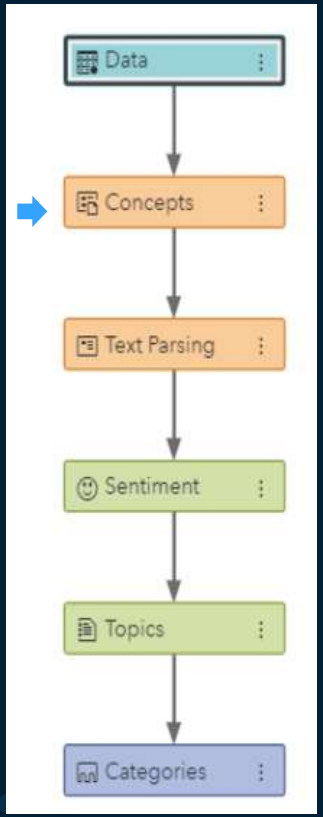
Copyright © SAS Institute Inc. All rights reserved.

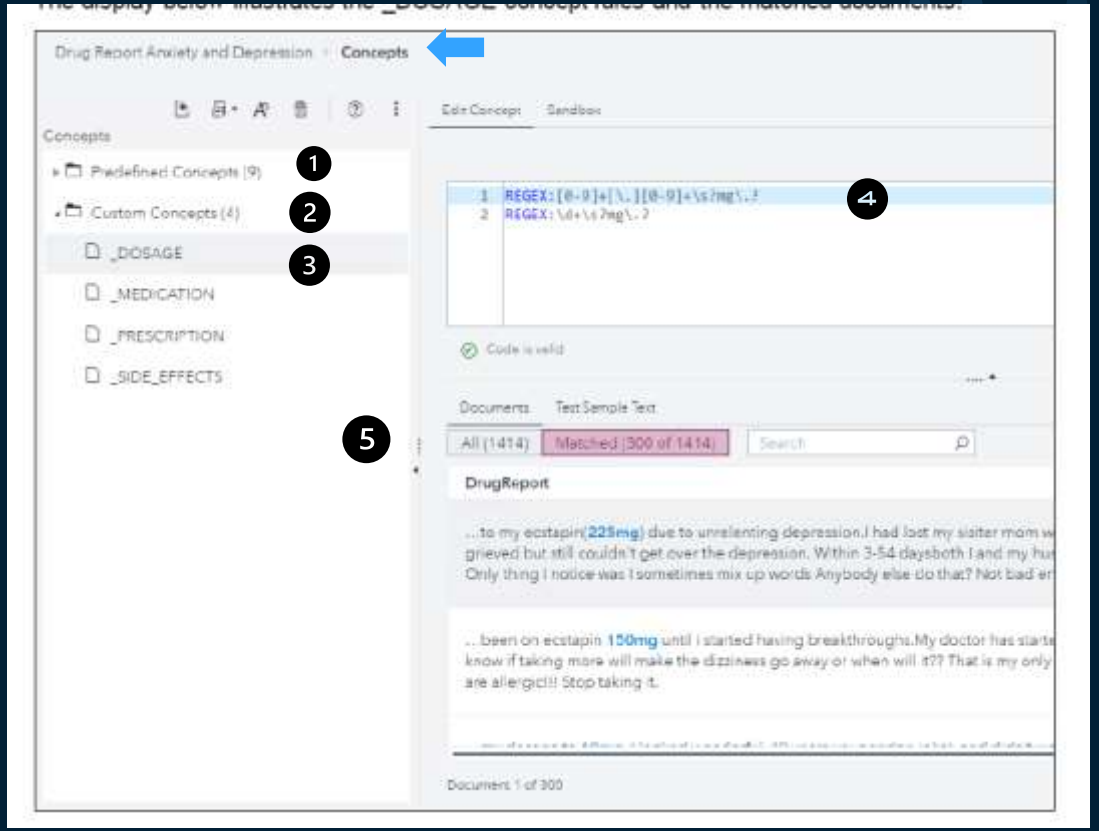
SAS Visual Text Analytics Project creation



Copyright © SAS Institute Inc. All rights reserved.

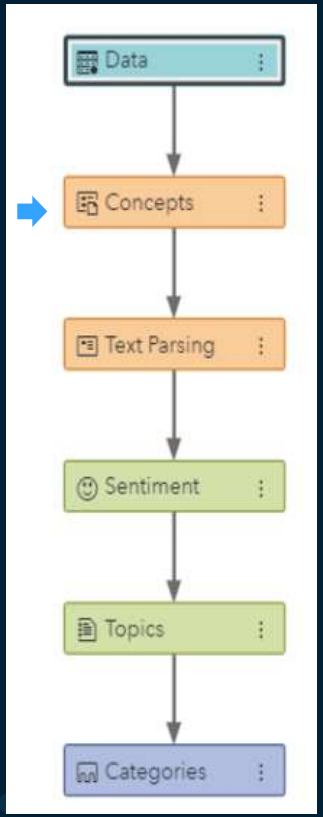
SAS Visual Text Analytics Steps – Concept extraction

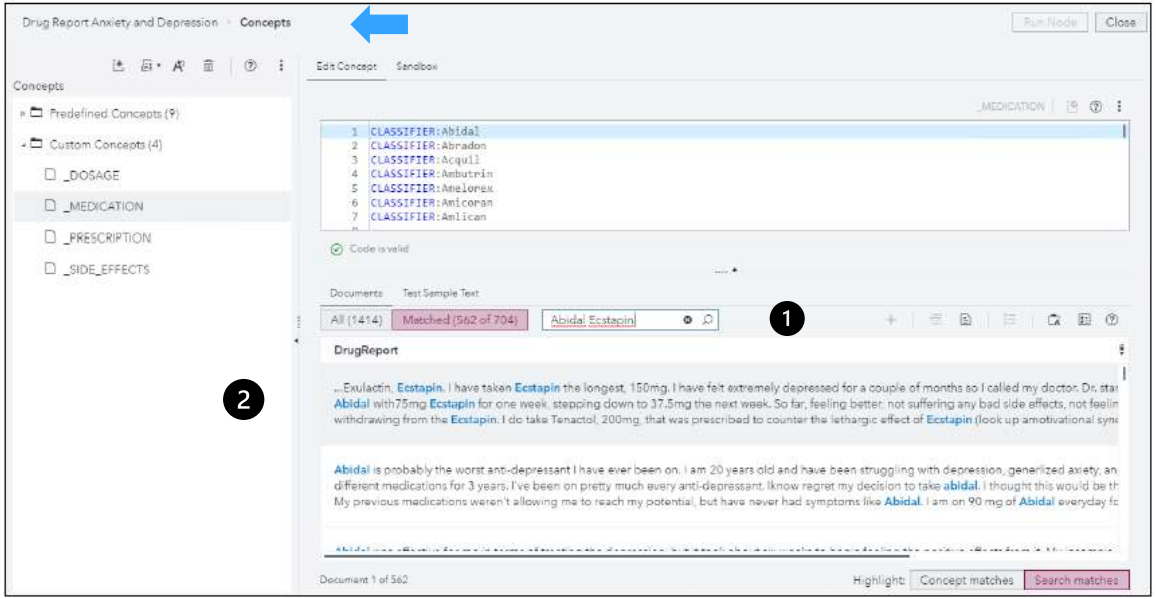




Copyright © SAS Institute Inc. All rights reserved.

Open Concept Node

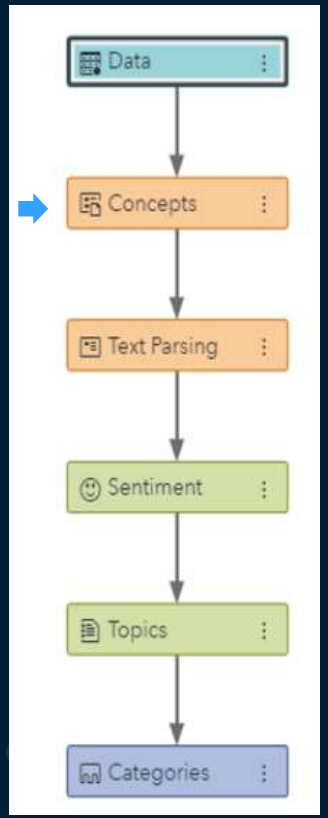




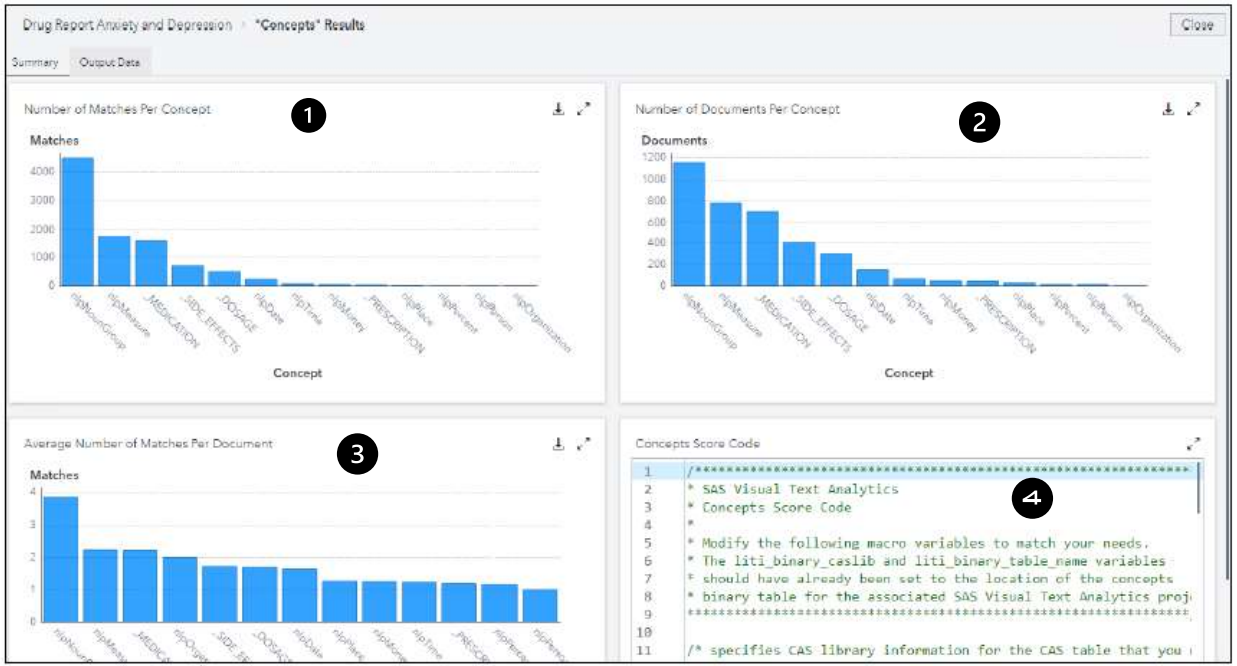
You see that 562 out of 704 documents matched the terms *Abidal* or *Ecstapin*.

Copyright © SAS Institute Inc. All rights reserved.

Concept node results

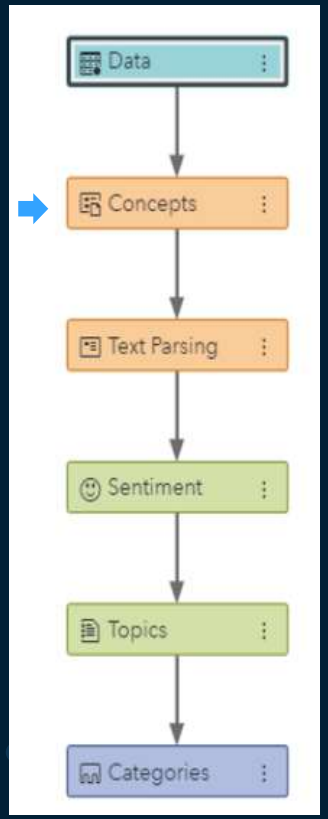


Select the **Concept** node and click **Results** option to see the results.



The Concept results window displays three bar charts: Number of Matches Per Concept, Number of Documents Per Concept, and Average Number of Matches Per Document.

Copyright © SAS Institute Inc. All rights reserved.



Click **Output Data** (to the right of the Summary at the upper left). Select **Output Tables** ⇨ **Transactional** ⇨ **View Output Data**.

Drug Report Anxiety and Depression › "Concepts" Results

Close

Summary | Output Data

1

2

3

Data sources

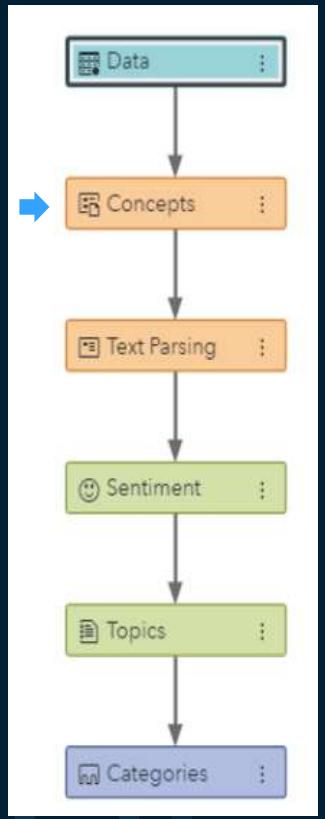
Output Tables

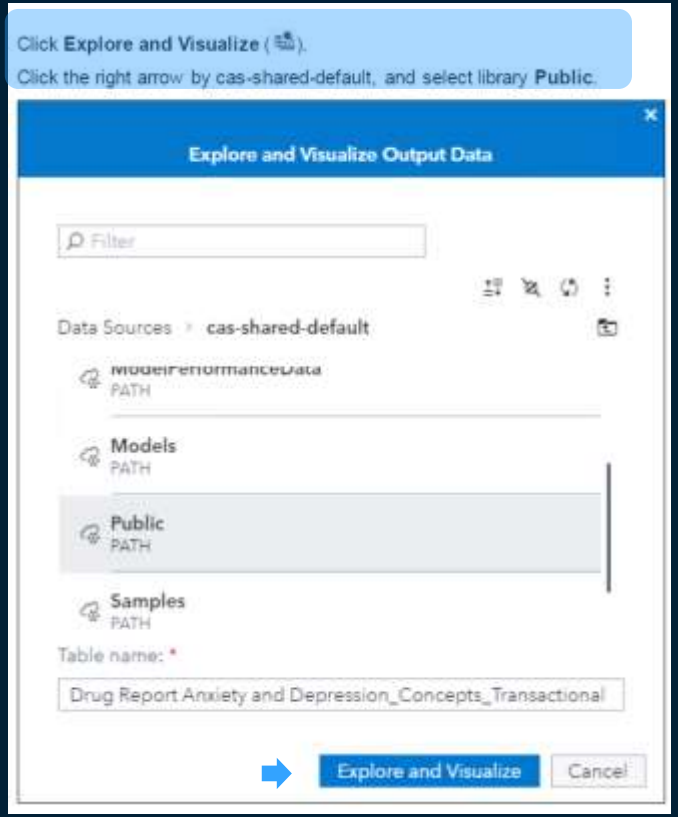
Transactional

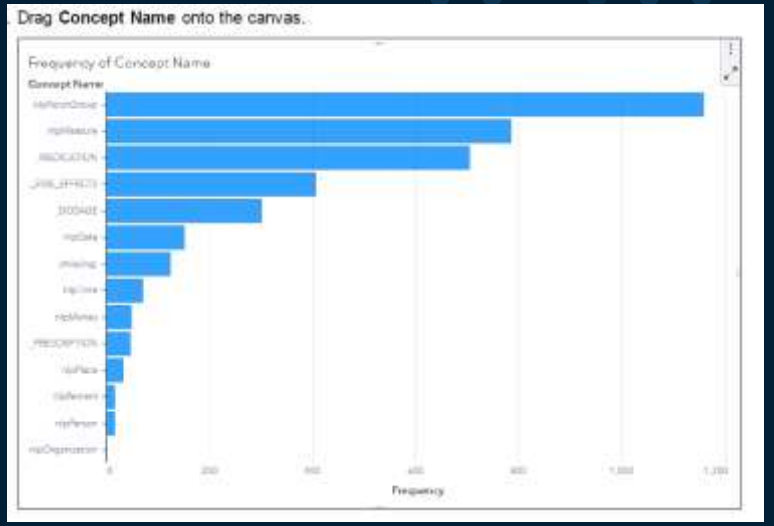
__uniqueid__	Text	Full Path	Concept Name	Keywords	Concept Type	ID
401	this med does perk you up, but after 4 weeks of	nlpMeasure	nlpMeasure	6-10 days | 4 weeks | 4-5 days | a month	Concept	0004
401	this med does perk you up, but after 4 weeks of	nlpNounGroup	nlpNounGroup	family support | cautious going | even though i |	Concept	0004
401	this med does perk you up, but after 4 weeks of	_MEDICATION	_MEDICATION	fortifex | prexifen	Concept	0004
902	I was on this for 9 months, up to 90mg a day. Had	_SIDE_EFFECTS	_SIDE_EFFECTS	dizziness	Concept	1094
902	I was on this for 9 months, up to 90mg a day. Had	_DOSAGE	_DOSAGE	90mg. | 30mg | 60mg | 60mg | 90mg | 30mg.	Concept	1094
902	I was on this for 9 months, up to 90mg a day. Had	nlpMeasure	nlpMeasure	30mg | 60mg | a day | 90mg | 3 days | 90mg | 9	Concept	1094
902	I was on this for 9 months, up to 90mg a day. Had	nlpNounGroup	nlpNounGroup	usual side | 60mg daily | slower rate | brain zaps | final	Concept	1094
902	I was on this for 9 months, up to 90mg a day. Had	nlpPlace	nlpPlace	My Dr.	Concept	1094
902	I was on this for 9 months, up to	_MEDICATION	_MEDICATION	Abidal	Concept	1094

Copyright © SAS Institute Inc. All rights reserved.

Exploring and visualize the scored data in SAS VA



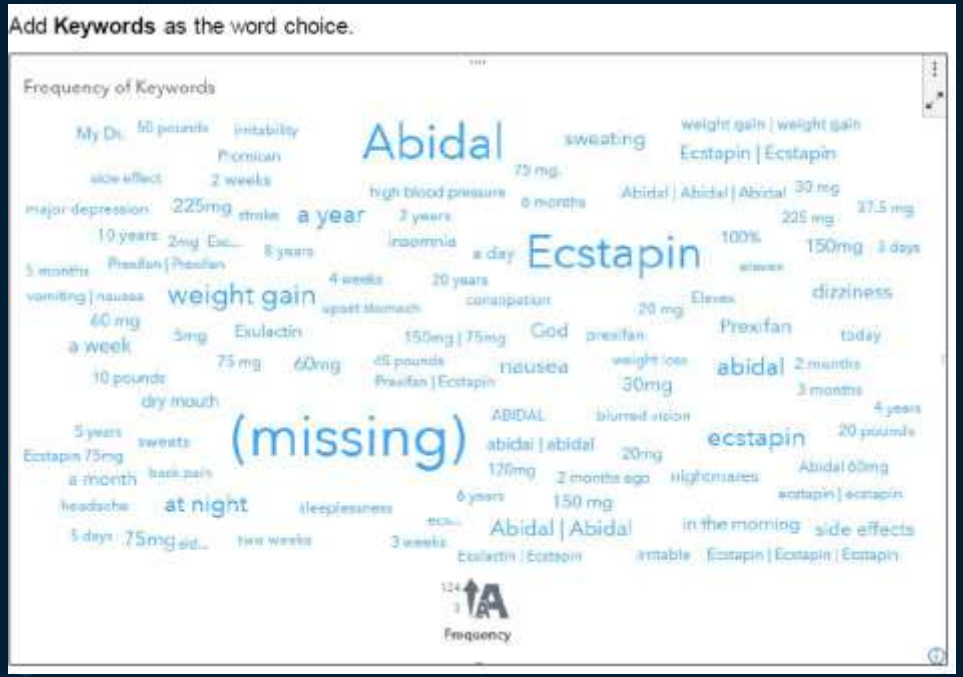




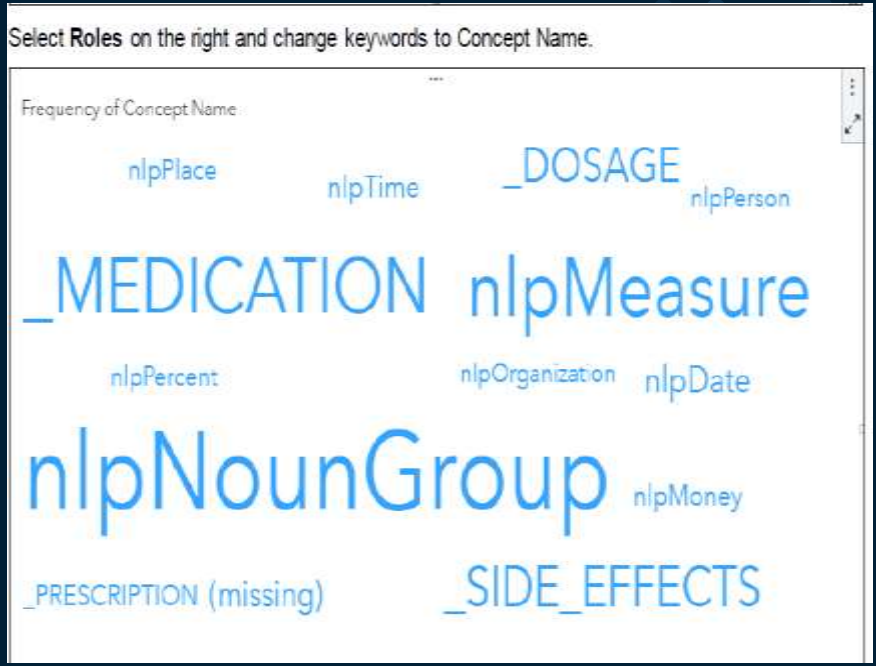
Copyright © SAS Institute Inc. All rights reserved.

Creating text cloud maps in VA

Add **Keywords** as the word choice.

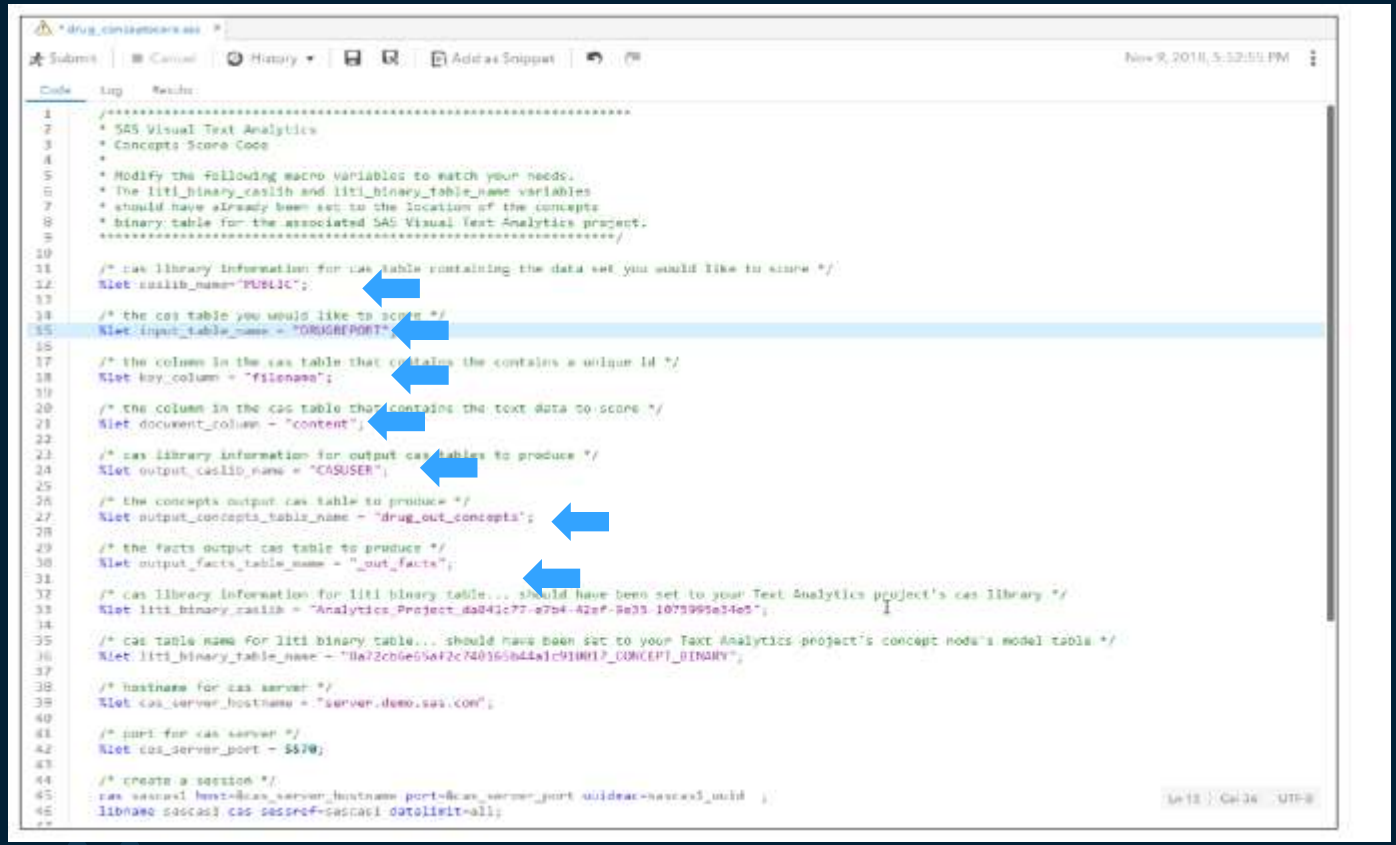


Select **Roles** on the right and change keywords to Concept Name.



Copyright © SAS Institute Inc. All rights reserved.

Scoring concept models



Copyright © SAS Institute Inc. All rights reserved.

SAS Visual Text Analytics: Term Extraction

Parse defined concepts

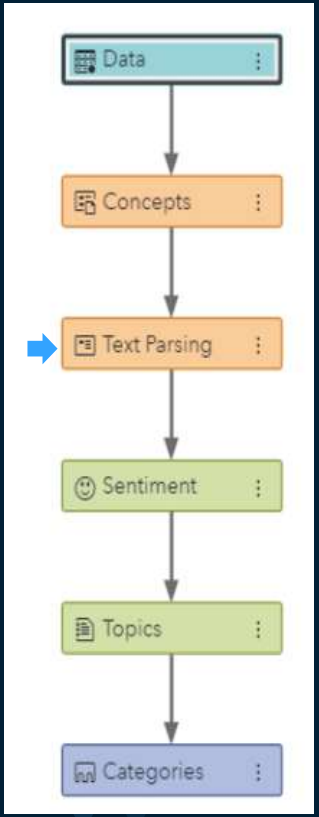
General language elements

NLP Tools

- parsing
- tokenization
- part-of-speech (POS) tagging
- synonym detection
- stemming

Copyright © SAS Institute Inc. All rights reserved.

Text parsing output



. Select the **Text Parsing** node and select **Open**.

Drug Report Anxiety and Depression › Text Parsing - Manage Terms

Run Node | Close

Kept Terms (1511) | Filter

Term	Role	Documents	Frequency ↓
not	ADV	658	1174
take	V	676	1104
depression	N	492	616
year	N	436	561
feel	V	379	541
work	V	379	487
drug	N	342	486
day	N	344	479

Dropped Terms (6711) | Filter

Term	Role	Documents	Frequency ↓
i	PRO	1081	5003
be	V	1127	4155
have	V	1069	2849
and	CONJ	1028	2747
the	DET	903	2558
to	PPOS	895	2383
it	PRO	878	2081
a	DET	792	1655

Documents

All (1414) | Matched | Search

DrugReport

This medication made me gain 40 pounds it has been 2 years and I have only lost 10 pounds. Beware and watch your weight.

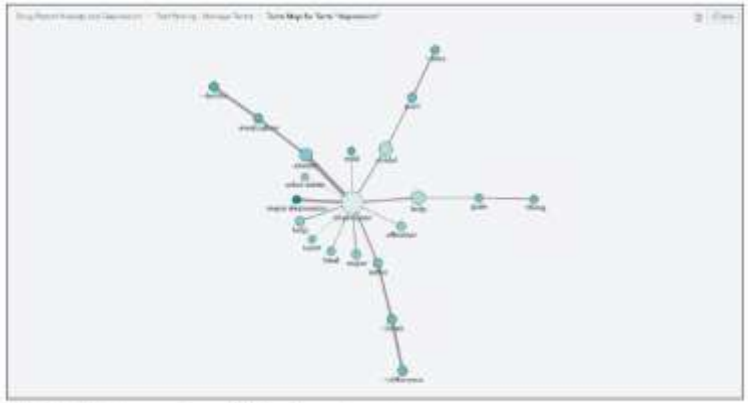
causing extreme anger, to the point that my family has become afraid of me. Doing things I would have considering doing before, like challanging the police, starting arguments, really, really wanting to beat the crap out of someone. Enough to the point that it scares me. But my doctor at this time refuses to take me off this medication. I am scared I will end up in jail for severly hurting someone.

Document 1 of 1414

Copyright © SAS Institute Inc. All rights reserved.

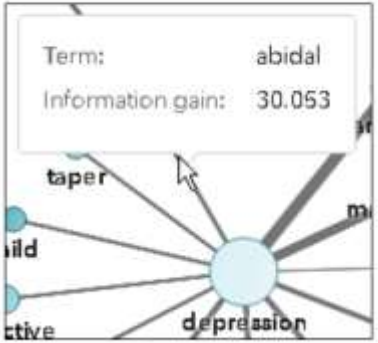
Term map in the text parsing node

Select the kept term **depression**. Click (**Term map**).



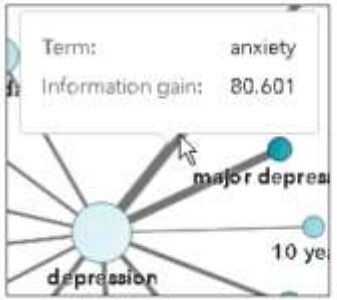
A total of 492 documents contain the term *depression*.

The line that connects *depression* and *abidal* has a thickness that is relative to the information gain (IG). You can right-click the line to display the IG value.



A total of 292 documents contain the drug name *abidal*, and 142 of those *abidal* documents also contain the word *depression*.

The largest information gain is associated with *anxiety*.



Copyright © SAS Institute Inc. All rights reserved.

Sentiment Analysis

use proprietary rules
to imply sentiment

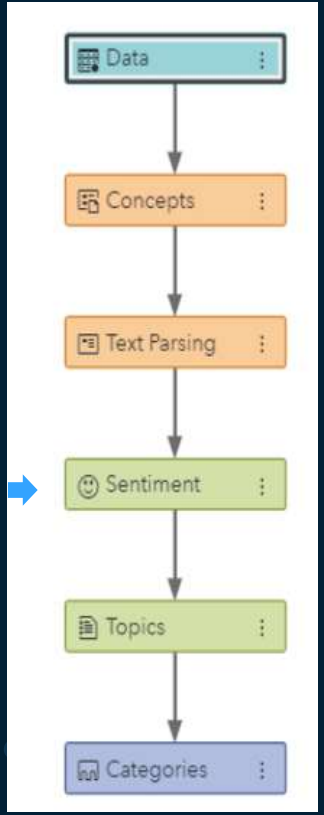
accommodate user-created
sentiment models

assign sentiment score to reflect
positive, neutral, or negative

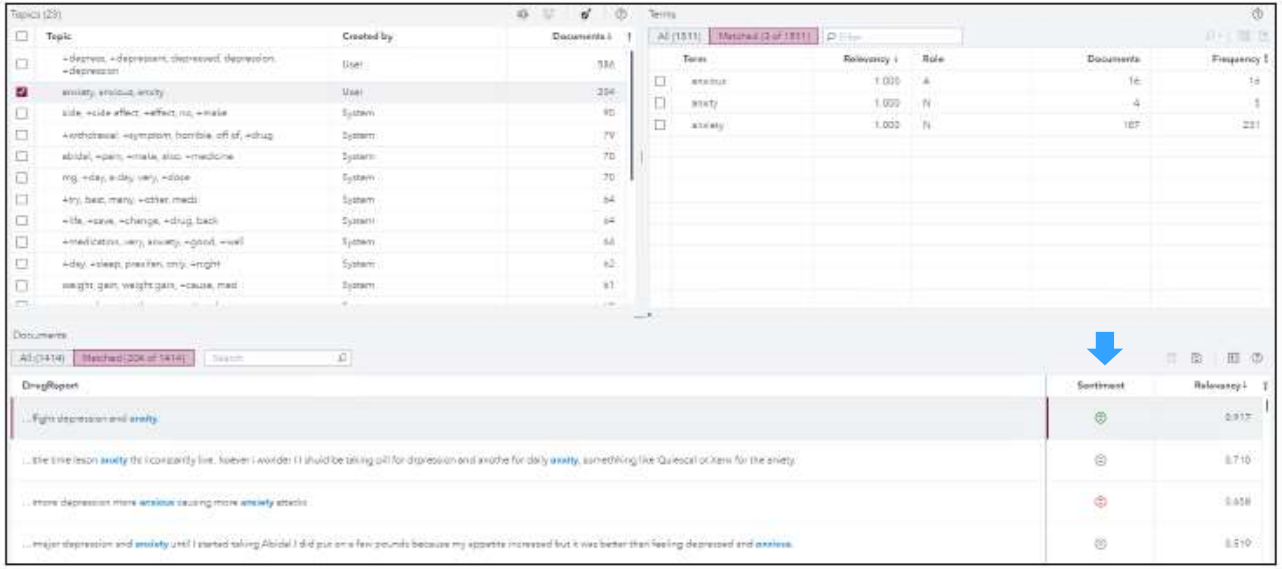
transfer sentiment scores
to successor nodes

Copyright © SAS Institute Inc. All rights reserved.

System generated sentiment scores



The new anxiety topic identifies 204 documents.

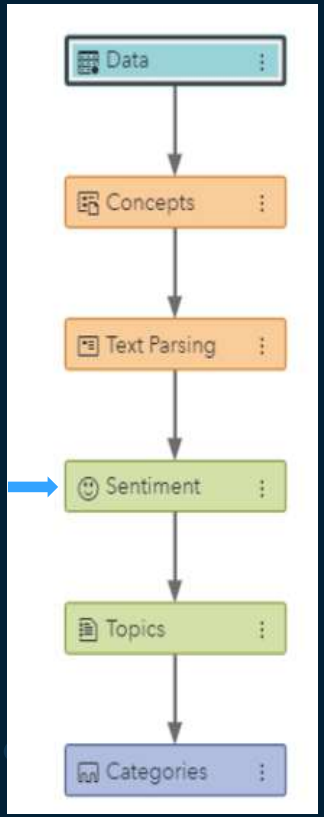


Finally, a user-defined, combined depression-and-anxiety-based topic is created by merging the three depression-based and anxiety-based user-defined topics. Select the three User topics and click **Merge topics**. A fourth User topic is created.

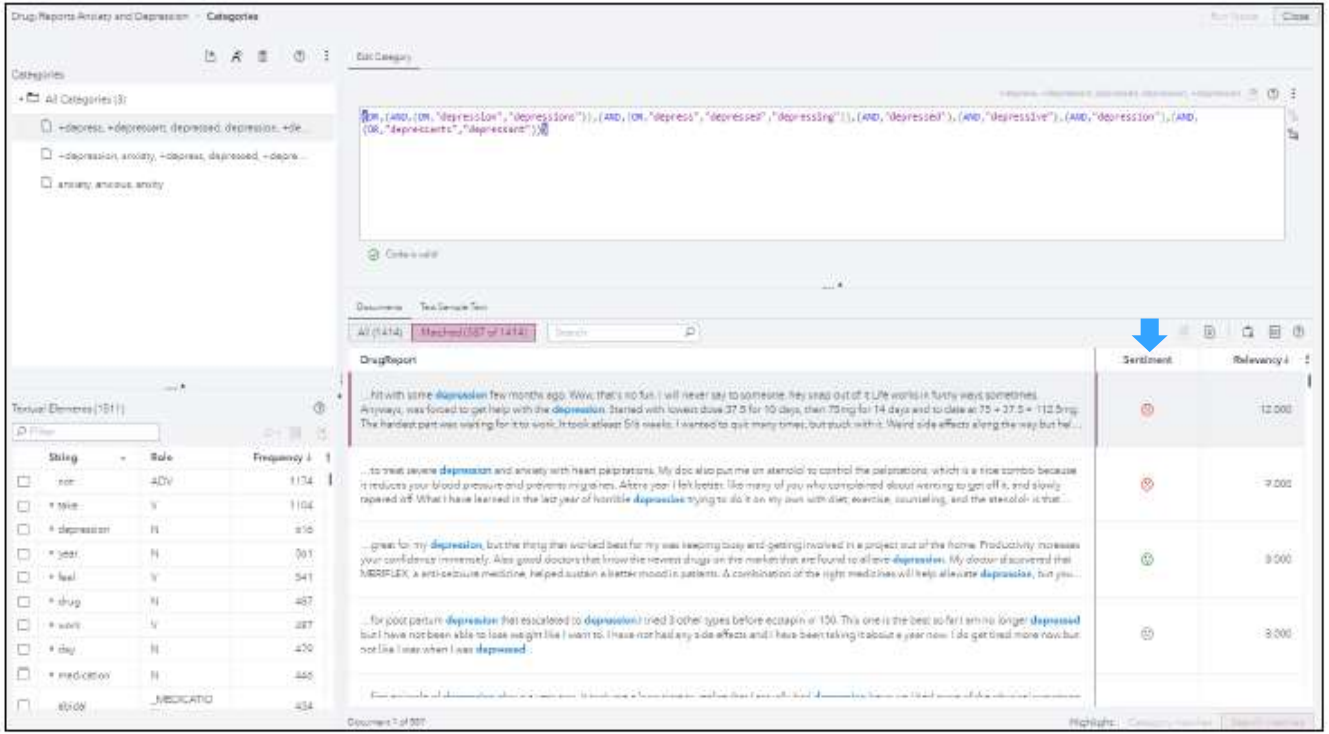
Click **Run Node**.

Copyright © SAS Institute Inc. All rights reserved.

System generated sentiment scores

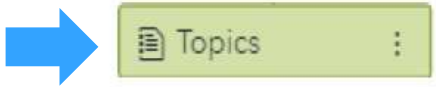


Examine each promoted topic category rules.



Copyright © SAS Institute Inc. All rights reserved.

Automatically Discover Hidden Topics: Machine Learning



automatically discovers core topics

relevant documents ranked per topic

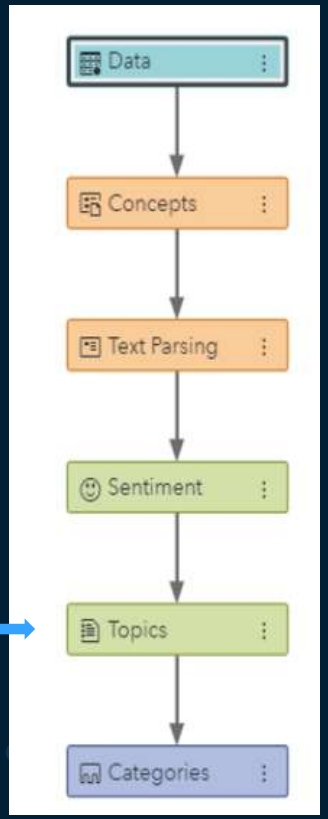
topics merged or split

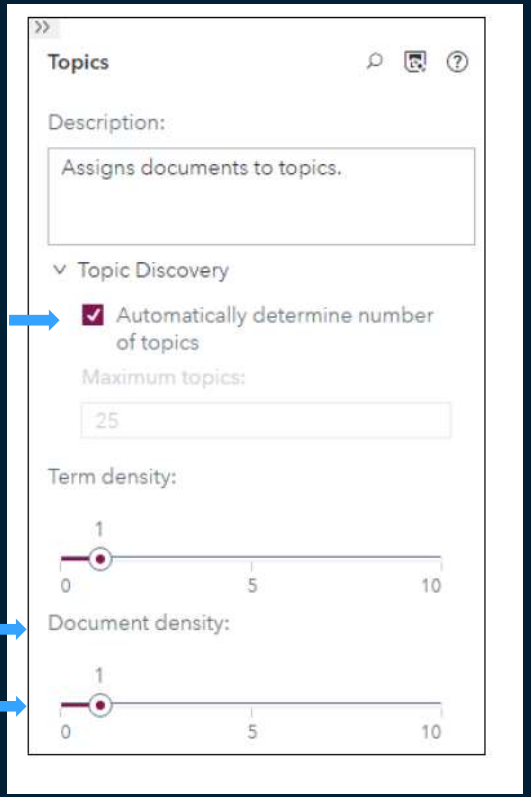
document-level sentiment included

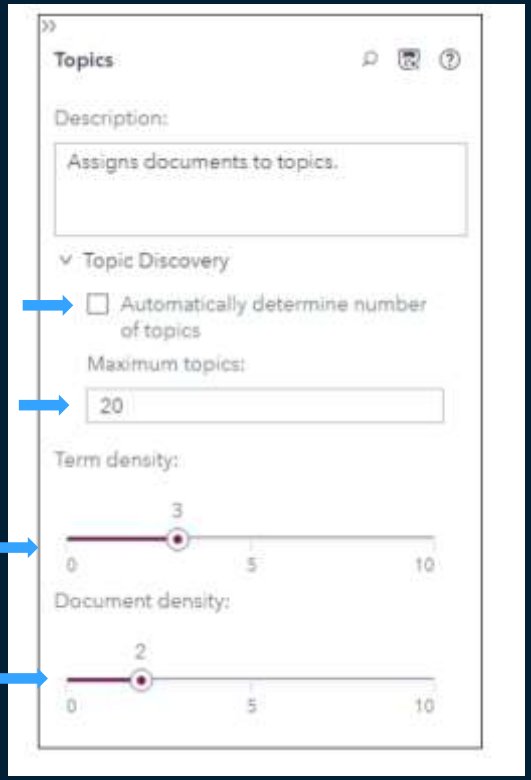
derived topics modified to create custom

Copyright © SAS Institute Inc. All rights reserved.

Modifying Topics discovery settings

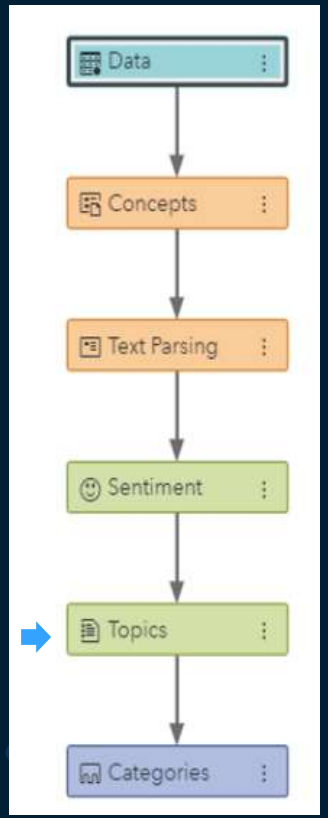






Copyright © SAS Institute Inc. All rights reserved.

System generated Topics

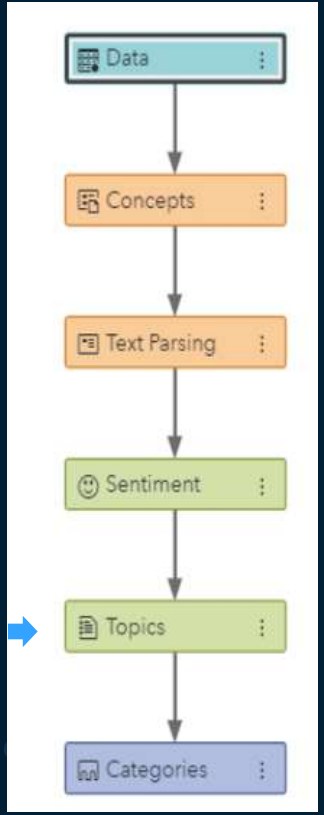


Topics (20)

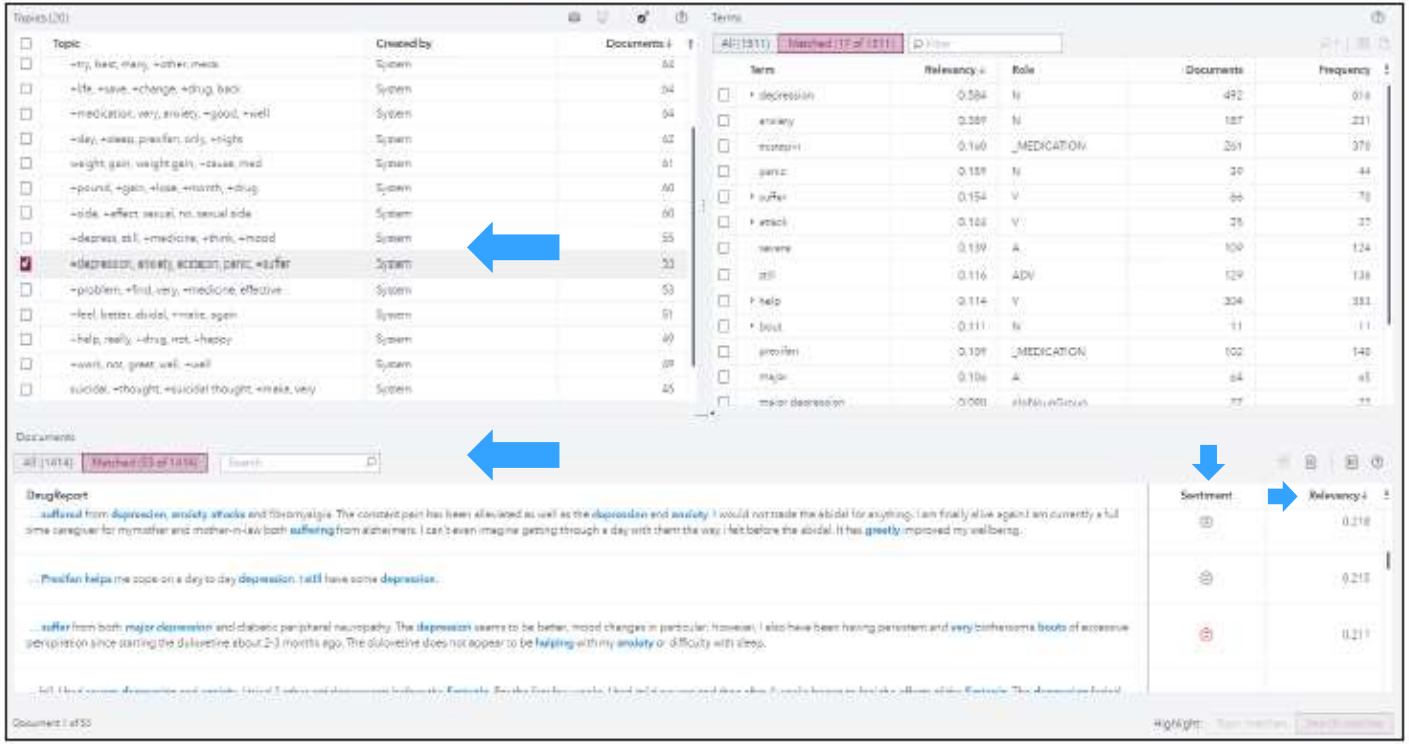
Topic	Created by	Documents
side, +side effect, +effect, no, +make	System	90
+withdrawal, +symptom, horrible, off of, +drug	System	79
abidal, +pain, +make, also, +medicine	System	70
mg, +day, a day, very, +dose	System	70
+try, best, many, +other, meds	System	64
+life, +save, +change, +drug, back	System	64
+medication, very, anxiety, +good, +well	System	64
+day, +sleep, prexifan, only, +night	System	62
weight, gain, weight gain, +cause, med	System	61
+pound, +gain, +lose, +month, +drug	System	60
+side, +effect, sexual, no, sexual side	System	60
+depress, still, +medicine, +think, +mood	System	55
+depression, anxiety, ecstapin, panic, +suffer	System	53
+problem, +find, very, +medicine, effective	System	53
+feel, better, abidal, +make, again	System	51
+help, really, +drug, not, +happy	System	49
+work, not, great, well, +well	System	49
suicidal, +thought, +suicidal thought, +make, very	System	45
+start, just, +week, med, +know	System	39
blood, pressure, blood pressure, anxiety, +high	System	35

Copyright © SAS Institute Inc. All rights reserved.

Features of system generated topics

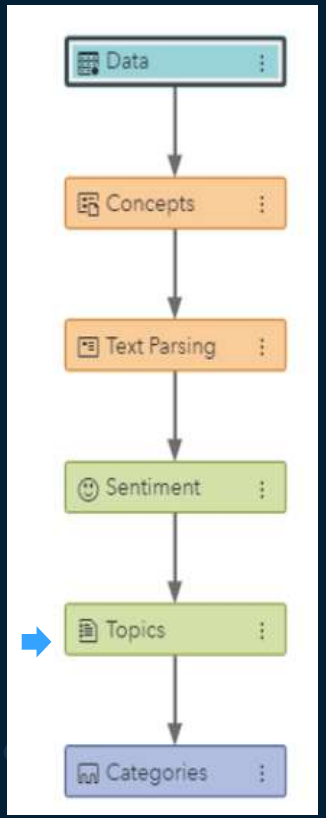


Select **+depression, anxiety, ecstapin, panic, +suffer**.



Copyright © SAS Institute Inc. All rights reserved.

Creating user defined topics by merging other topics



Increasing the maximum number of topics to 20 resulted in a new depression-based topic (**+depress, still, +medicine, +think, +mood**) with 55 matched documents and 22 matched terms. Thus, increasing the number of maximum topics did not help identify more depression- and anxiety-related documents.

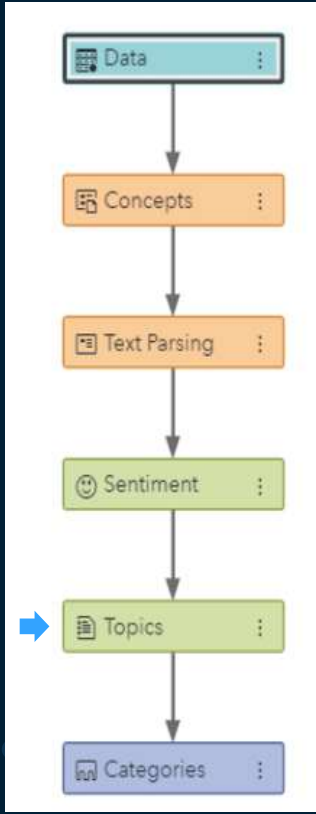
- Select the two depression topics and click **Merge topics**. A new topic is added as a User (rather than System) topic.
- Click **Run Node**.

Topics (21)

Topic	Created by	Documents ↓
side, +side effect, +effect, no, +make	System	90
+withdrawal, +symptom, horrible, off of, +drug	System	79
abidal, +pain, +make, also, +medicine	System	70
mg, +day, a day, very, +dose	System	70
+try, best, many, +other, meds	System	64
+life, +save, +change, +drug, back	System	64
+medication, very, anxiety, +good, +well	System	64
+day, +sleep, prexifen, only, +night	System	62
weight gain, weight gain, +cause, med	System	61
+pound, +gain, +lose, +month, +drug	System	60
+side, +effect, sexual, no, sexual side	System	60
+depression, still, anxiety, +depress, ecstapin	User	57

Copyright © SAS Institute Inc. All rights reserved.

Creating a custom topic from terms



Deselect all topics. In the term pane, click the **All** tab, and sort by **Term**. Select all terms that start with *depress*. There should be 10 terms selected.

Click (**Create a topic from terms**). This action creates a second User topic. There are now 20 System topics and two User topics.

Click **Run Node**.

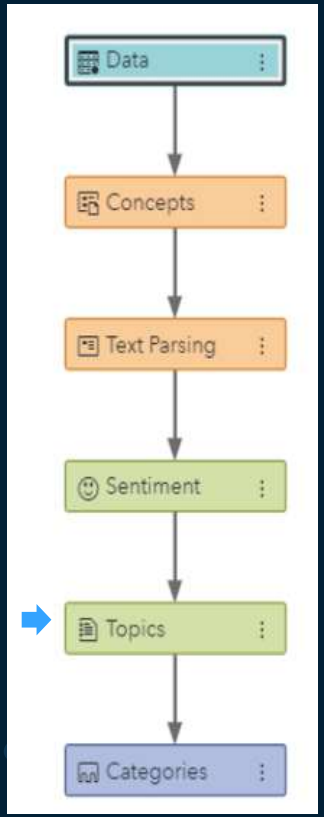
The new topic captures 586 documents.

Topics (22)

Topic	Created by	Documents ↓
+depress, +depressant, depressed, depression, +depression	User	586
side, +side effect, +effect, no, +make	System	90
+withdrawal, +symptom, horrible, off of, +drug	System	79
abidal, +pain, +make, also, +medicine	System	70
mg, +day, a day, very, +dose	System	70
+try, best, many, +other, meds	System	64
+life, +save, +change, +drug, back	System	64
+medication, very, anxiety, +good, +well	System	64
+day, +sleep, prexifan, only, +night	System	62
weight, gain, weight gain, +cause, med	System	61
+pound, +gain, +lose, +month, +drug	System	60

Copyright © SAS Institute Inc. All rights reserved.

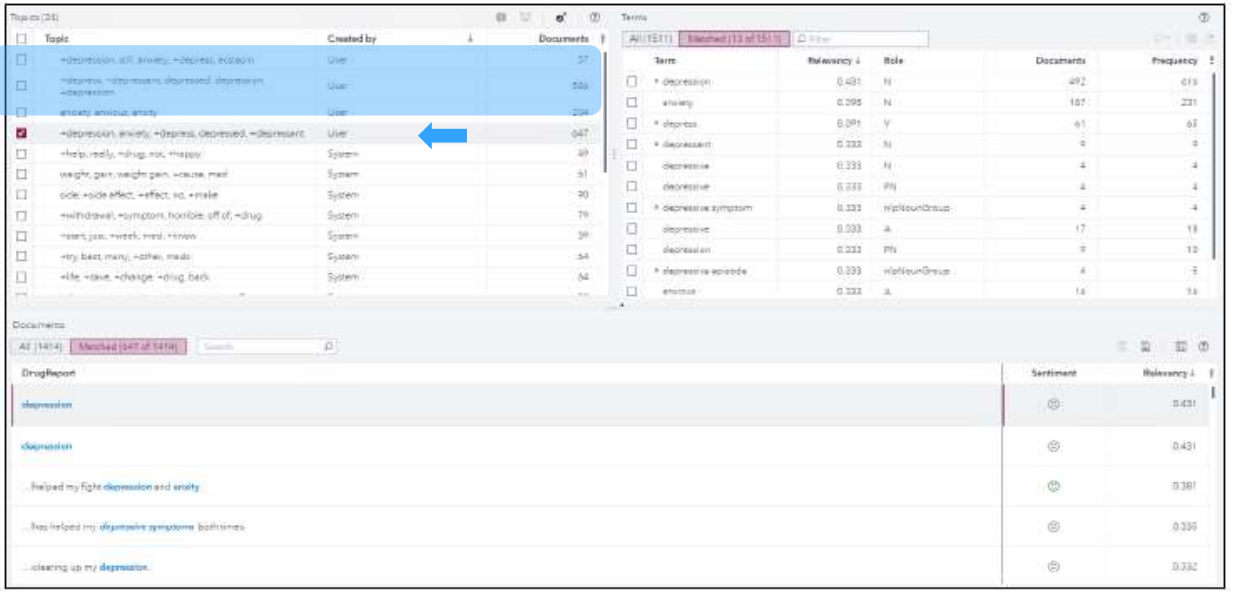
Creating user defined topics by merging user defined topics



Finally, a user-defined, combined depression-and-anxiety-based topic is created by merging the three depression-based and anxiety-based user-defined topics. Select the three User topics and click **Merge topics**. A fourth User topic is created.

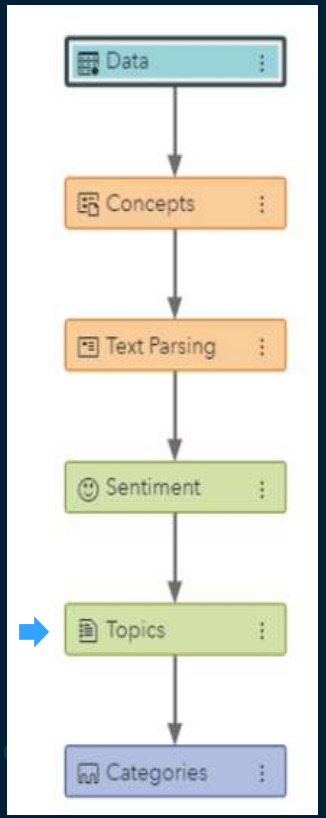
Click **Run Node**.

The newest User topic identifies 647 documents.



Copyright © SAS Institute Inc. All rights reserved.

Creating another custom topic



Add another User topic by selecting all terms related to anxiety. The terms **anxiety**, **anxious**, and **anxity** are selected.

Terms

All (1511) | Matched | Filter

	Term	Role	Documents	Frequency
☐	anti-depressant	N	51	57
☐	antidepressant	N	52	59
☑	anxiety	N	187	231
☑	anxious	A	16	16
☑	anxity	N	4	5
☐	anymore	ADV	45	47
☐	anything	N	34	37
☐	anyway	ADV	7	8
☐	anywhere	ADV	5	6
☐	appetite	N	20	20

Table

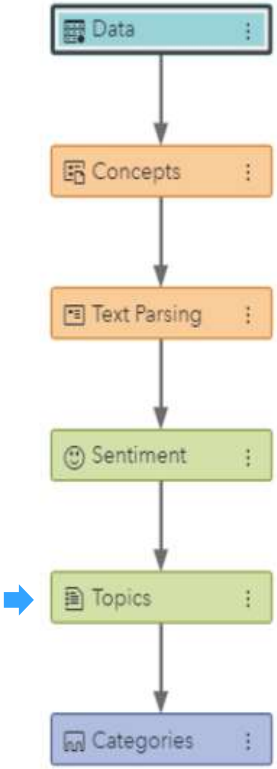
Description automatically generated

Click the icon for **Create a topic from terms**.

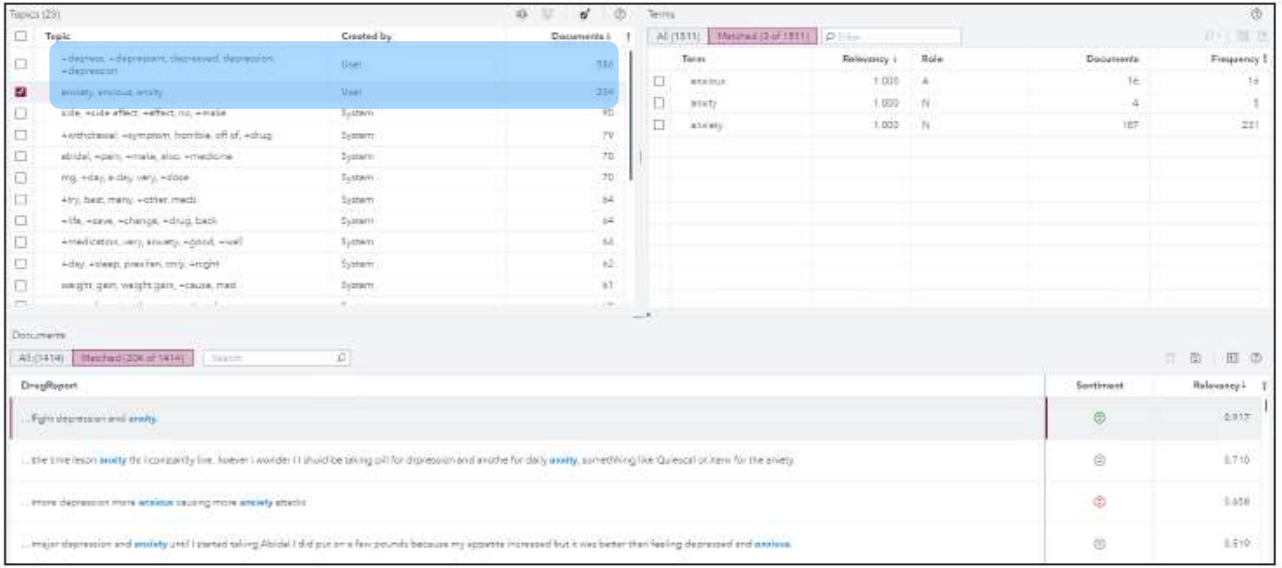
Click **Run Node**.

Copyright © SAS Institute Inc. All rights reserved.

Creating another custom topic



The new anxiety topic identifies 204 documents.



Finally, a user-defined, combined depression-and-anxiety-based topic is created by merging the three depression-based and anxiety-based user-defined topics. Select the three User topics and click **Merge topics**. A fourth User topic is created.

Click **Run Node**.

Copyright © SAS Institute Inc. All rights reserved.

A sample Topic scored data file

The output table **ASRS_out_documents** is created and stored in the PUBLIC library.

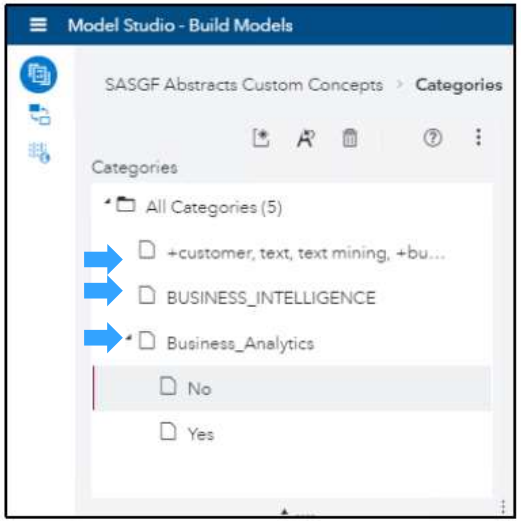
PUBLIC.ASRS_OUT_DOCUMENTS | Columns: 32 of 32 | Total rows: 21519 | Rows 1 to 200

Enter expression

	Col1	_Col2_	_Col3_	_Col4_	_Col5_	_Col6_	_Col7_	_Col8_	_Col9_	_Col1...
1	0.19847354 63	0.00643179 44	0.08643419 02	0.19534131 19	0.01030578 2	0.00670471 97	0.09100688 65	0.07881529 22	-0.0113816 8	-0.0394757 02
2	0.01001343 45	0.01193305 31	0.01314013 28	0.00962410 33	0.07854393 97	0.01444503 21	0.00485029 94	0.01410577 68	0.00628118 96	0.01111935 87
3	0.05846577 64	0.00494122 73	0.04873458 16	0.04605954 64	-0.0144930 23	0.04485330 82	0.06051552 53	0.00377150 36	0.08580722 04	0.05197937 26
4	-0.0111955 26	0.01486958 18	0.10565853 77	0.05208292 53	0.05097773 61	0.00826112 82	0.00465248 86	-0.0156777 25	0.00489825 24	0.11555344 99
5	0.04617909 26	-0.0362552 86	0.02473544 15	0.00207705 79	-0.0326571 68	-0.0227175 62	0.54798575 16	0.25292977 28	0.04759192 15	-0.0414928 31
6	0.02979279 32	0.02245643 63	0.05213735 1	0.02768965 32	0.12076104 34	0.21113382 08	0.01544792 62	0.02200903 29	0.01810194 79	-0.0100048 02
7	0.02125490 8	0.07013831 4	0.00213814 42	0.06251029 28	-0.0301033 07	0.00207380 85	0.03137579 06	-0.0249775 06	-0.0039019 03	0.01432084 82
8	-0.0317092 51	0.02660648 25	0.02948603 27	0.16134114 11	0.00498050 61	0.02108605 43	0.03394550 17	-0.0264215 38	0.00283823 92	0.01201520 99
9	0.04669131 47	0.01753994 66	0.02462550 27	0.03384969 97	-0.0041408 24	0.00132597 49	0.18782173 35	0.14231479 71	0.01314466 53	0.02318901 54
10	0.00145617 33	0.15270084 39	0.01704378 97	0.04772194 2	0.00941408 56	0.03286804 83	0.04361390 18	0.00340108 39	-0.0048414 81	0.04136030 5
11	0.07526212 51	0.00256392 7	0.08370938 69	0.23151136 25	-0.0480711 39	-0.0290987 75	0.30566481 59	0.03134610 72	0.03438437 91	-0.0329465 74
12	0.05113851 08	0.05279136 14	0.02084400 18	0.09924056 69	0.03714441 85	0.02160616 21	0.05749636 91	0.06281262 14	0.02495957 03	0.00738071 77

Copyright © SAS Institute Inc. All rights reserved.

Categorization Rule Generation



Boolean linguistic rules automatically derived

easy to understand

can be edited

Copyright © SAS Institute Inc. All rights reserved.

Category Rules

Boolean Rules

CATEGORY

- Boolean and proximity operators
- arguments
- modifiers

- ✓ more precise than linguistic rules
- ✓ term located for structured input documents (for example, XML)
- ✓ qualifiers modify the effect of the rule

Copyright © SAS Institute Inc. All rights reserved.

Category rules

Boolean Rule Syntax

- Boolean operator names are specified in uppercase.
- Boolean operators and their arguments are enclosed in parentheses.
- Arguments are separated by commas.
- Terms are enclosed in double quotation marks.

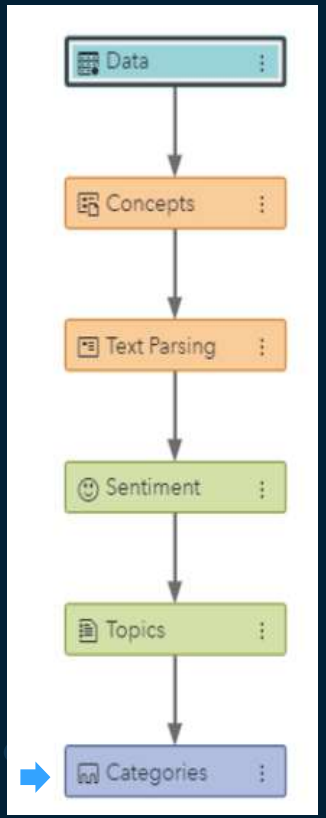
Example, when displayed in text view:

```
(AND,"Amazon",(OR,"online retailer","e-commerce"))
```

This rule is true when the term *Amazon* appears with either of the terms *online retailer* or *e-commerce*.

Copyright © SAS Institute Inc. All rights reserved.

Categories are generated from topics



Select the three User topics at the top based on the sort order of the number of documents.

Topics (24)

	Topic	Created by	Documents
☐	+depression, still, anxiety, +depress, ecstapin	User	57
☑	+depress, +depressant, depressed, depression, +depression	User	586
☑	anxiety, anxious, anxity	User	204
☑	+depression, anxiety, +depress, depressed, +depressant	User	647
☐	+help, really, +drug, not, +happy	System	49
☐	weight, gain, weight gain, +cause, med	System	61
☐	side, +side effect, +effect, no, +make	System	90
☐	+withdrawal, +symptom, horrible, off of, +drug	System	79

Select **Add topics as categories**.

Topics (24)

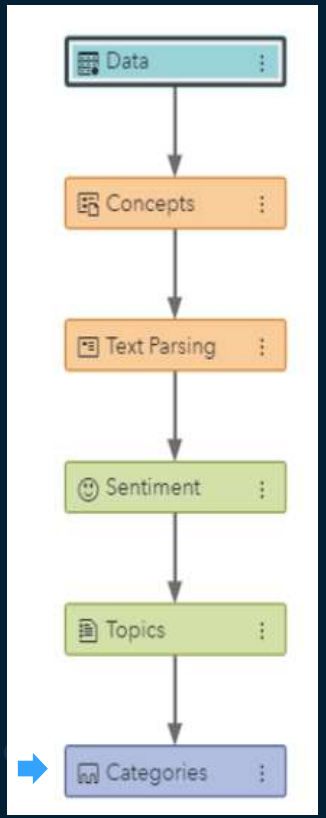
	Topic	Created by	Documents
☐	+depression, still, anxiety, +depress, ecstapin	User	57
☑	+depress, +depressant, depressed, depression, +depression	User	586
☑	anxiety, anxious, anxity	User	204
☑	+depression, anxiety, +depress, depressed, +depressant	User	647
☐	+help, really, +drug, not, +happy	System	49
☐	weight, gain, weight gain, +cause, med	System	61
☐	side, +side effect, +effect, no, +make	System	90
☐	+withdrawal, +symptom, horrible, off of, +drug	System	79

Close the **Topics** node.

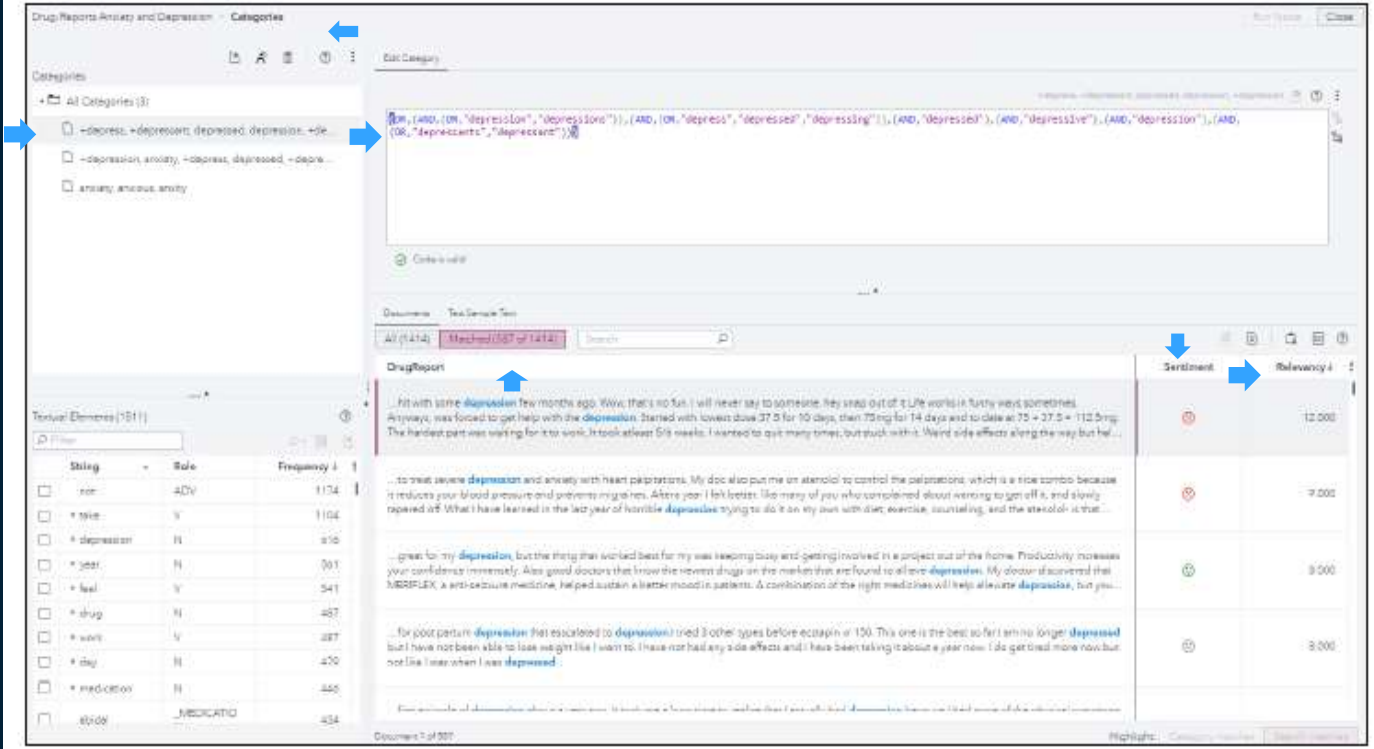
Run the pipeline.

Copyright © SAS Institute Inc. All rights reserved.

Category rules

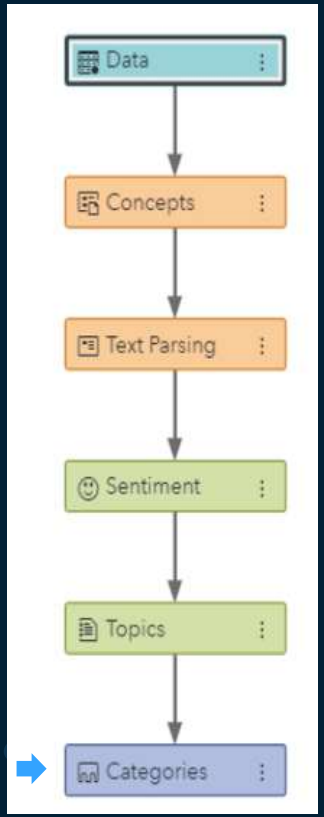


Examine each promoted topic category rules.



Copyright © SAS Institute Inc. All rights reserved.

Scoring Category model



. Upload the **ScoreCode_Categories_NYC_Low.sas** file to the server and open the file in the SAS Studio editor. Make the following changes to the score code:

```
%let input_caslib_name = "PUBLIC";
%let input_table_name = "NYCDOHMH_A";
%let key_column = "RECORD_ID";
%let document_column = "VIOLATION_DESCRIPTION";
%let output_caslib_name = "PUBLIC";
%let output_categories_table_name = "nyc_out_categories";
%let output_matches_table_name = "nyc_out_matches";
%let output_modeling_ready_table_name = "nyc_out_modeling_ready";
```

. Run the program. Three output tables are written to the PUBLIC library.

Code Log Results

The CAS Procedure
textRuleScore.applyCategory
Output CAS Tables

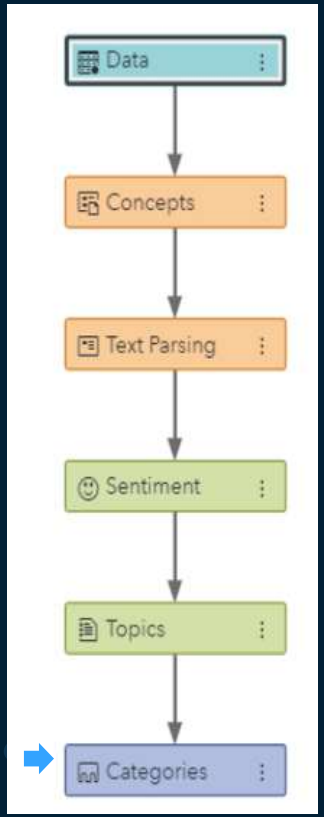
Results from textRuleScore.applyCategory

Output CAS Tables

CAS Library	Name	Label	Number of Rows	Number of Columns
Public	nyc_out_categories		56298	4
Public	nyc_out_matches		206241	5
Public	nyc_out_modeling_ready		20530	12

Copyright © SAS Institute Inc. All rights reserved.

Scored Output data from category node



Variable Name	Label	Type	Role	Level
category_1	_VerminCategory	Numeric	Input	Binary
category_10	_FoodPrepCategory	Numeric	Input	Binary
category_11	_FacilitiesCategory	Numeric	Input	Binary
category_2	_DocumentationCategory	Numeric	Input	Binary
category_3	SCOREGRADE/C	Numeric	Input	Binary
category_4	_WorkersCategory	Numeric	Input	Binary
category_5	vermin, +facility, present, live, 's	Numeric	Input	Binary
category_6	SCOREGRADE/A	Numeric	Input	Binary
category_7	SCOREGRADE/B	Numeric	Input	Binary
category_8	LowGrade/0	Numeric	Input	Binary
category_9	LowGrade/1	Numeric	Input	Binary
LowGrade	Low Grade Indicator	Numeric	Target	Binary
RECORD_ID	Document ID	Character	ID	Nominal
SCORE		Numeric	Rejected	Interval
SCOREGRADE	Score Grade	Character	Rejected	Nominal
VIOLATION_DESCRIPTION		Character	Rejected	Nominal

Copyright © SAS Institute Inc. All rights reserved.

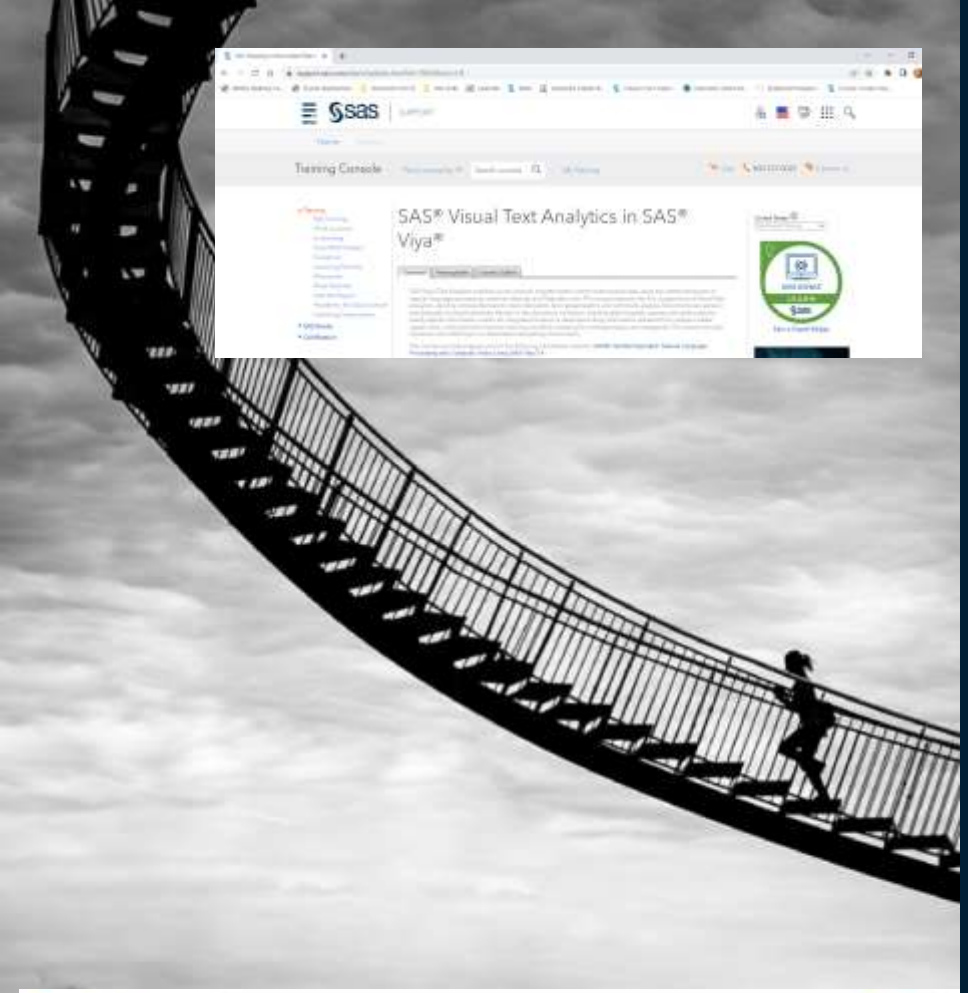
Summary

Introduced Visual Text Analytics in SAS Viya

Text analytics steps

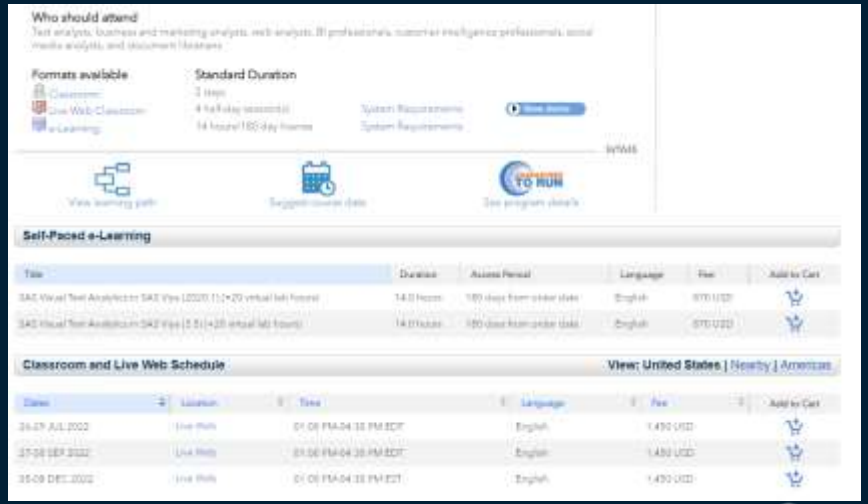
Scoring new text data for text features

Copyright © SAS Institute Inc. All rights reserved.



Thank you

Questions / Comments



Inc. All rights reserved.

Q&A

Please submit your questions using the Q&A icon located in the menu at the bottom of your screen

Copyright © SAS Institute Inc. All rights reserved.

Explore Helpful Resources

Ask the Expert
View other user webinars that provide insights into using SAS products to make your job easier.

FREE Training
Learn from home – free for 30 days. Get software labs to practice and online support if needed.

SAS Support Communities
Ask questions, get answers and share insights with SAS users.

SAS Analytics Explorers
An exclusive platform to collaborate, learn and share your expertise. Gain access to a diverse network to advance your career. Special rewards and recognition exclusively for SAS users.

SAS Users YouTube Channel
A plethora of videos on hundreds of topics, just for SAS users.

Newsletters
Get the latest SAS news plus tips, tricks and more.

Users Groups
Meet local SAS users, network and exchange ideas – virtually.

SAS Profile
If you haven't already done so, create your SAS Profile to access free training, SAS Support Communities, technical support, software downloads, newsletters and more.

Copyright © SAS Institute Inc. All rights reserved.

Thank you

for joining us for this SAS webinar

Copyright © SAS Institute Inc. All rights reserved.