Paper SAS3349-2019

Automatic LITI Rule Generation for Information Extraction

Teresa Jade, Xu Yang, Christina Hsiao, and Aaron Arthur, SAS Institute Inc.

ABSTRACT

Have you ever wished that you could get assistance with writing rules in LITI (language interpretation for textual information)? Now you can! Information extraction in SAS® Visual Text Analytics uses the LITI syntax, which is a powerful way to uncover useful information hidden in textual data. The extracted information can be used as new structured data for predictive models, for faceted searches, or can be added to reports to enable better business decisions. In this presentation and paper, we demonstrate how to use the rule generation feature in SAS Visual Text Analytics. We also explain how the rules are generated in the background and what you can expect from the resulting rule set. The resulting rules can be used directly in the model, edited by the user, or used in an exploratory way to find new candidate rule elements. We provide tips and best practices for using rule generation to build your information extraction models faster and easier.

INTRODUCTION

LITI rules are a powerful tool for Information Extraction. They can be simple or complex, depending on the type of model you need for your business case. Predefined rule sets (concepts) are one way to use LITI without writing any of your own rules, because these concepts are pre-built to find common types of pieces of information like person names, place names, organization names, dates, times, money amounts, and percentages. Another pre-built rule set in Visual Text Analytics is called nlpNounGroup. This concept is a grammatical pattern that identifies additional multiple-word terms that can be useful for topic identification or for LITI rules. In this paper, we describe automatic rule generation — a new way to get pre-built rules in the LITI syntax — and show you some ways to leverage this feature.

There are three ways to generate rules in Visual Text Analytics using automatic rule generation. First, you can simply highlight text in the documents and select the command to create a CLASSIFIER rule. This type of rule matches literal strings of text. Second, fact rule generation automatically creates 1-2 PREDICATE_RULE rules from two concepts. This type of rule identifies relationships between two regular concepts, (for example, a symptom such as "redness" and the body part it manifests on such as "arm"). Third, concept rule generation creates up to 25 rules based on the rules you have "seeded" in the concept, either manually or using the first method of creating CLASSIFIER rules.

The business problem we examine in this paper leverages data from publicly available descriptions of side effects or adverse events that patients have reported following a vaccination. This Vaccine Adverse Event Reporting System (VAERS) is managed by the CDC and FDA. Among other objectives, these agencies use the system as follows:

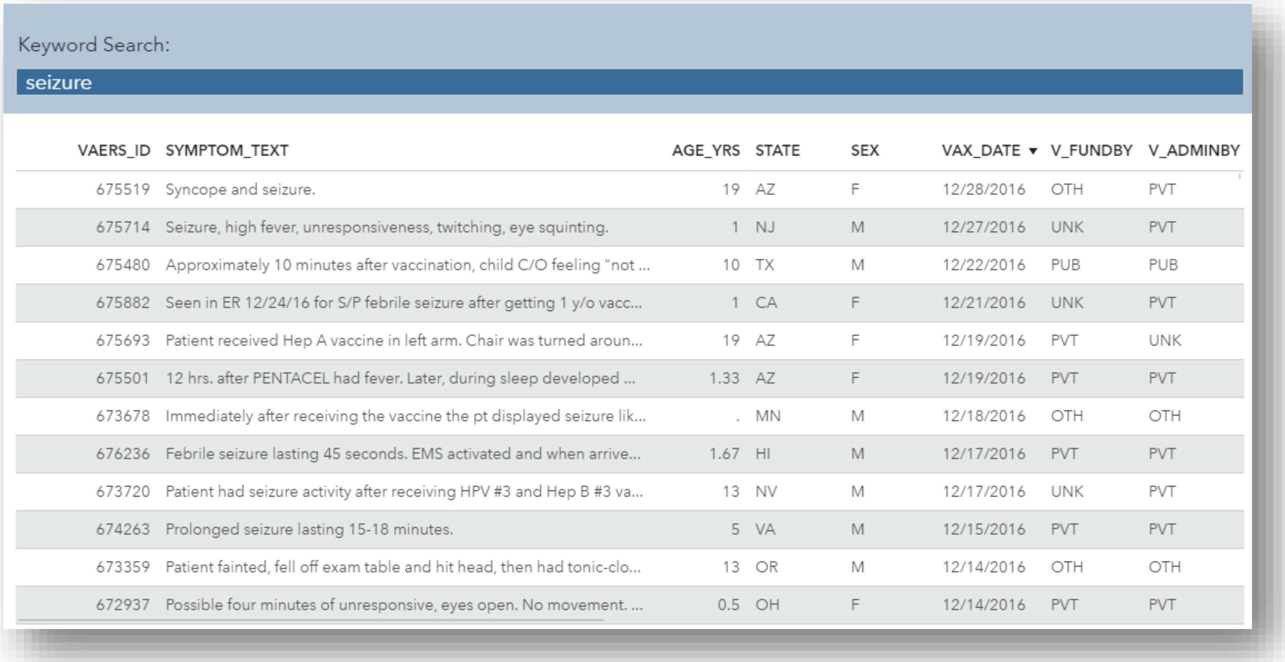
- To monitor increases in known adverse events and detect new or unusual vaccine adverse events
- To identify potential patient risk factors, including temporal, demographic, or geographic reporting clusters

This data has the symptoms listed in a separate table from the narrative text that describes the patient experience. For the purposes of this paper, we will pretend that the symptoms table does not exist and extract that information from the narrative text contained in a variable called SYMPTOM_TEXT. If you were an analyst working at the CDC or FDA, you might find that the symptoms table is inaccurate or wonder if it is missing information. Using information extraction is one way to check your information by comparing it with another data field. One the other hand, if your data is like the single table used in our example and you do not already have the information you need in another table or column, this example demonstrates how to get that information out of the unstructured text data.

THE PROBLEM OF IDENTIFICATION OF SYMPTOMS

A dashboard from Visual Analytics shows the raw data in Figure 1. Not all the data is shown, but you can see some examples of the narrative in the SYMPTOM_TEXT column as well as other related information per patient, such as age, gender, U.S. state, vaccination date, and where the funding and the vaccination came from.

Figure 1 Dashboard from SAS Visual Analytics



Keyword Search:

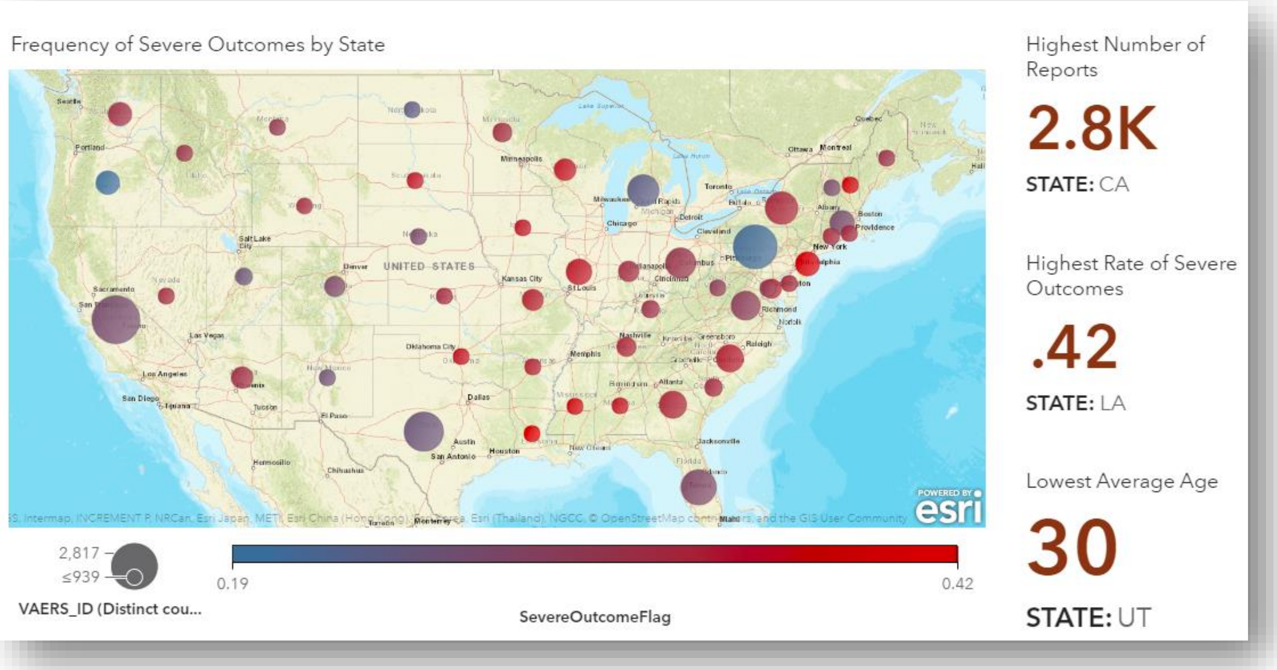
seizure

VAERS_ID	SYMPTOM_TEXT	AGE_YRS	STATE	SEX	VAX_DATE ▼	V_FUNDBY	V_ADMINBY
675519	Syncope and seizure.	19	AZ	F	12/28/2016	OTH	PVT
675714	Seizure, high fever, unresponsiveness, twitching, eye squinting.	1	NJ	M	12/27/2016	UNK	PVT
675480	Approximately 10 minutes after vaccination, child C/O feeling "not ...	10	TX	M	12/22/2016	PUB	PUB
675882	Seen in ER 12/24/16 for S/P febrile seizure after getting 1 y/o vacc...	1	CA	F	12/21/2016	UNK	PVT
675693	Patient received Hep A vaccine in left arm. Chair was turned aroun...	19	AZ	F	12/19/2016	PVT	UNK
675501	12 hrs. after PENTACEL had fever. Later, during sleep developed ...	1.33	AZ	F	12/19/2016	PVT	PVT
673678	Immediately after receiving the vaccine the pt displayed seizure lik...	.	MN	M	12/18/2016	OTH	OTH
676236	Febrile seizure lasting 45 seconds. EMS activated and when arrive...	1.67	HI	M	12/17/2016	PVT	PVT
673720	Patient had seizure activity after receiving HPV #3 and Hep B #3 va...	13	NV	M	12/17/2016	UNK	PVT
674263	Prolonged seizure lasting 15-18 minutes.	5	VA	M	12/15/2016	PVT	PVT
673359	Patient fainted, fell off exam table and hit head, then had tonic-clo...	13	OR	M	12/14/2016	OTH	OTH
672937	Possible four minutes of unresponsive, eyes open. No movement. ...	0.5	OH	F	12/14/2016	PVT	PVT

In this dashboard, it is easy to do a search for a specific symptom such as "seizure", but analysts need to more than just search. They need to see what patterns are emerging in aggregate, so they can detect systemic safety or process issues. Also, a challenge with this data and the use of keyword search to explore it is that the symptoms of the patient are documented in a narrative format that might also contain elements of a patient's medical history or statements about symptoms the patient is not experiencing in relation to the adverse event. We demonstrate ways to extract the patient's current symptoms from the narrative text to use in reporting and monitoring patterns across all adverse events or adverse event from one state or area of the United States.

The types of reports that can be created in Visual Analytics with the raw data presented in the table in Figure 1 include summary reports across the geography as shown in Figure 2. Figure 2 visualizes the number of reports and characterizes severe outcomes by state across the United States and highlights the state with the highest number of VAERS reports and the state with the highest rate of severe outcomes.

Figure 2 Frequency of Severe Outcomes by State



Frequency of Severe Outcomes by State

Highest Number of Reports

2.8K

STATE: CA

Highest Rate of Severe Outcomes

.42

STATE: LA

Lowest Average Age

30

STATE: UT

PROJECT SETUP AND DESIGN

To start the process of information extraction, open a new project in Visual Text Analytics and select the pipeline "Text Analytics: Assisted Concept Rule Creation". You can use any pipeline that contains a Concepts node for the modeling we will demonstrate here. However, to leverage Textual Elements in the Concepts node, make sure your pipeline has a Parsing node before the Concepts node. Select SYMPTOM_TEXT as the text variable to set up the data, turn on misspelling detection in the Text Parsing node properties, and then turn on predefined concepts in Concepts node properties, and run the node as in Figure 3.

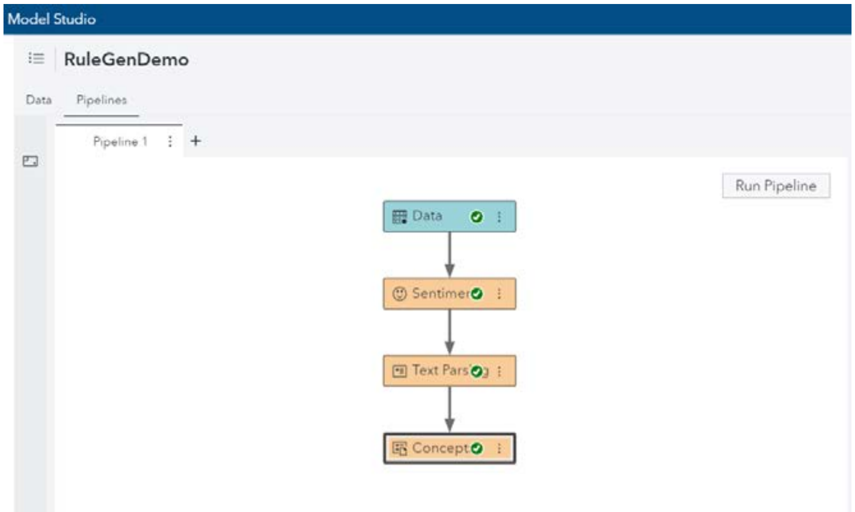


Figure 3 Pipeline Run for the First Time

For this paper, we will skip past the early exploratory stages of data analysis and go straight to building an information extraction model. During exploration, you had noticed that there are two types of symptoms you would want to model: symptoms that are of a general nature (body- or system-wide) or are localized by the term used to describe them, and symptoms that need to be specified with a body part to be understood. For example, the term "fever" is understood to be system-wide. No one will ask you where you have a fever. However, if you have a symptom like "redness", there will be questions to determine where on the body the symptom occurs. Redness on the face would potentially mean something different from redness on the upper arm.

To design this model, you designate and document four concepts as depicted in Figure 4. The top-level concept is the patientSymptom concept. This is the only concept that is primary and produces the final output from the model. The other concepts are all supporting concepts, but you don't mark them as supporting in the software until after development is completed, because otherwise you would not be able to see the test results of these concepts. The supporting concepts include: generalSymptom, localSymptom, and bodyPart.

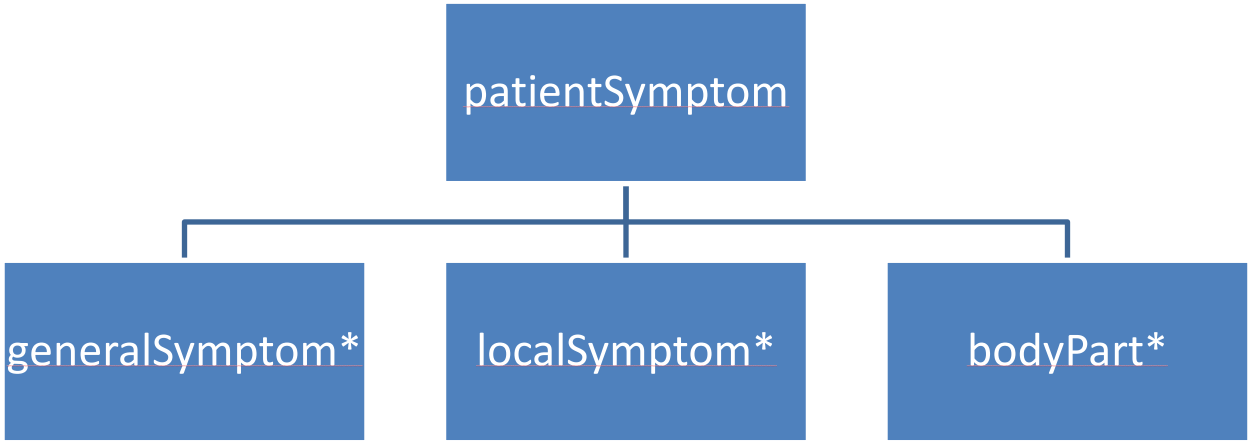
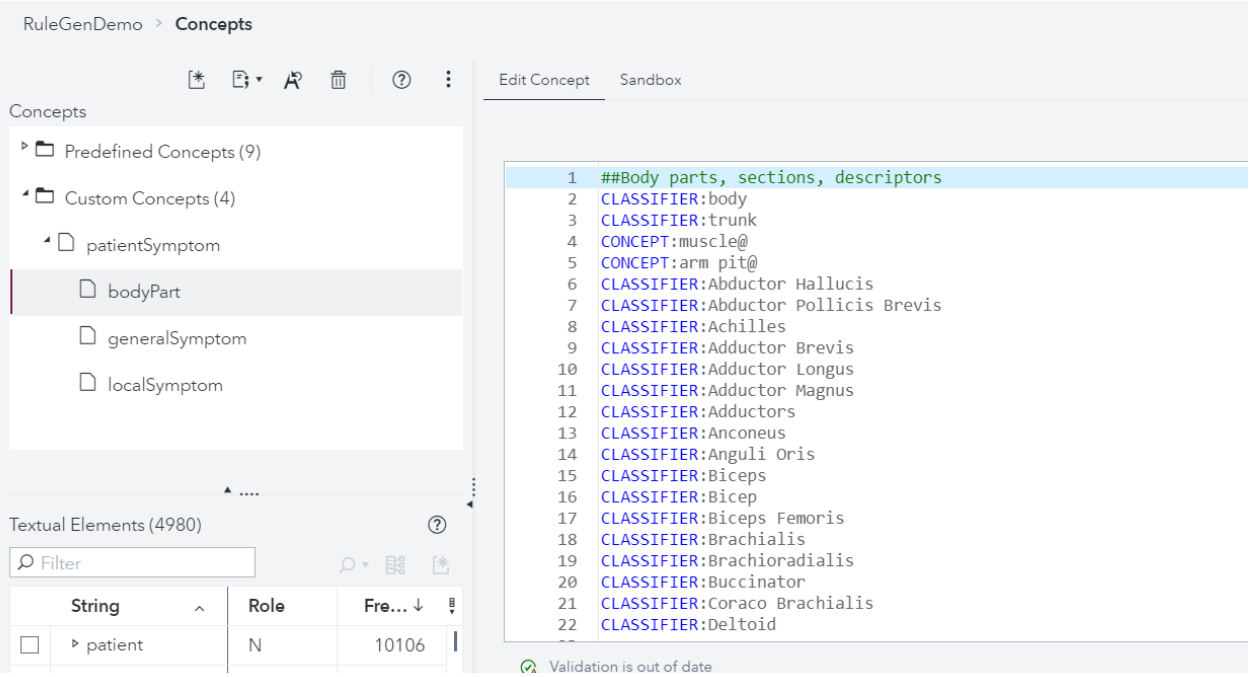


Figure 4 Design for Patient Symptom Model with Three Supporting Concepts

DEVELOPING BODY PART CONCEPT

Starting with the bodyPart concept, you first collect a list of parts from medical dictionaries and subject matter experts. This list is then prepared in LITI rule format and placed in the bodyPart node and you run the node to check that the initial rules, as shown in Figure 5, are performing as expected. You next examine the document matches, and see that the body parts are being matched as expected from these rules, but some body parts are not matched. One pattern that you notice in the data is that some body parts appear before words such as "left", "right", or "upper". So, one investigation you can do is to check to see if these contexts will lead to more candidate rules for body parts, by using some C_CONCEPT rules.

Figure 5 Initial Set of Body Part Rules

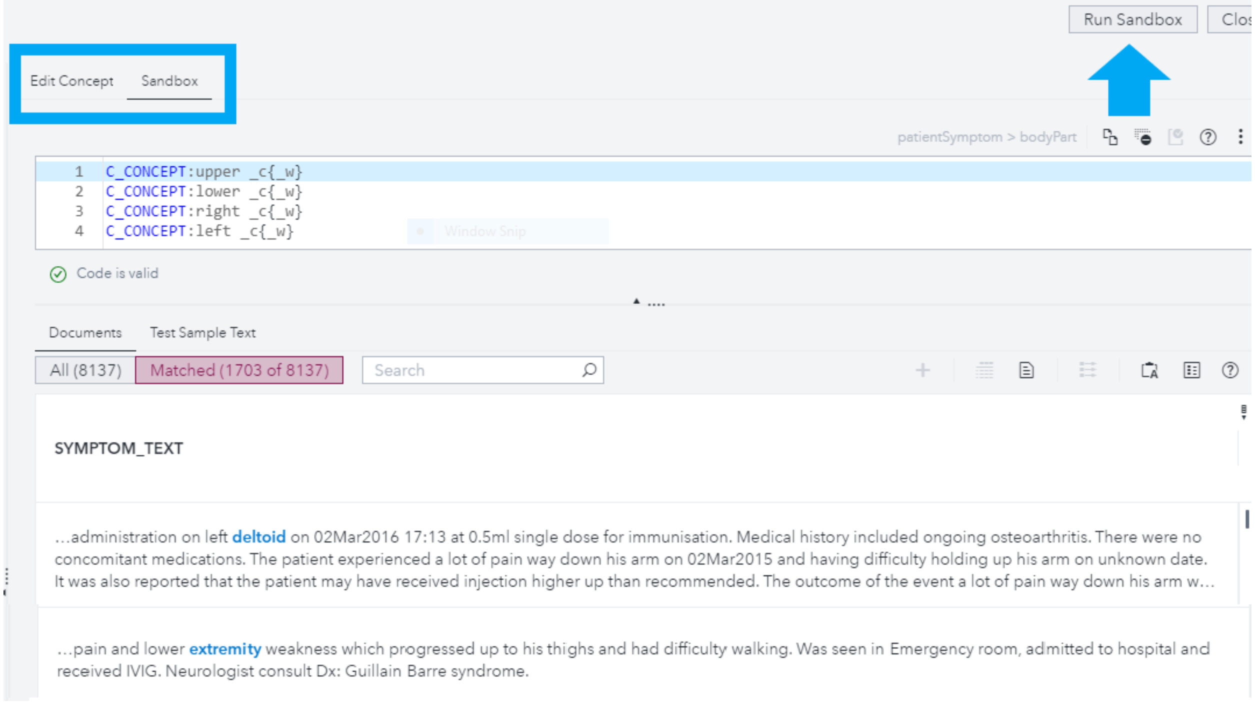


```
##Body parts, sections, descriptors
CLASSIFIER:body
CLASSIFIER:trunk
CONCEPT:muscle@
CONCEPT:arm pit@
CLASSIFIER:Abductor Hallucis
CLASSIFIER:Abductor Pollicis Brevis
CLASSIFIER:Achilles
CLASSIFIER:Adductor Brevis
CLASSIFIER:Adductor Longus
CLASSIFIER:Adductor Magnus
CLASSIFIER:Adductors
CLASSIFIER:Anconeus
CLASSIFIER:Anguli Oris
CLASSIFIER:Biceps
CLASSIFIER:Bicep
CLASSIFIER:Biceps Femoris
CLASSIFIER:Brachialis
CLASSIFIER:Brachioradialis
CLASSIFIER:Buccinator
CLASSIFIER:Coraco Brachialis
CLASSIFIER:Deltoid
```

SANDBOX INTRODUCTION

To explore the contexts, you can use the new Sandbox editor to test the exploratory rules quickly. When you work in the Sandbox (to the right of your regular rules editor – see Figure 6), you can test your concept and any concept that it references in a quicker test than one that uses the entire model.

Figure 6 Using the Sandbox Editor to Test bodyPart Rules Against All Documents



```
C_CONCEPT:upper _c{_w}
C_CONCEPT:lower _c{_w}
C_CONCEPT:right _c{_w}
C_CONCEPT:left _c{_w}
```

...administration on left **deltoid** on 02Mar2016 17:13 at 0.5ml single dose for immunisation. Medical history included ongoing osteoarthritis. There were no concomitant medications. The patient experienced a lot of pain way down his arm on 02Mar2015 and having difficulty holding up his arm on unknown date. It was also reported that the patient may have received injection higher up than recommended. The outcome of the event a lot of pain way down his arm w...

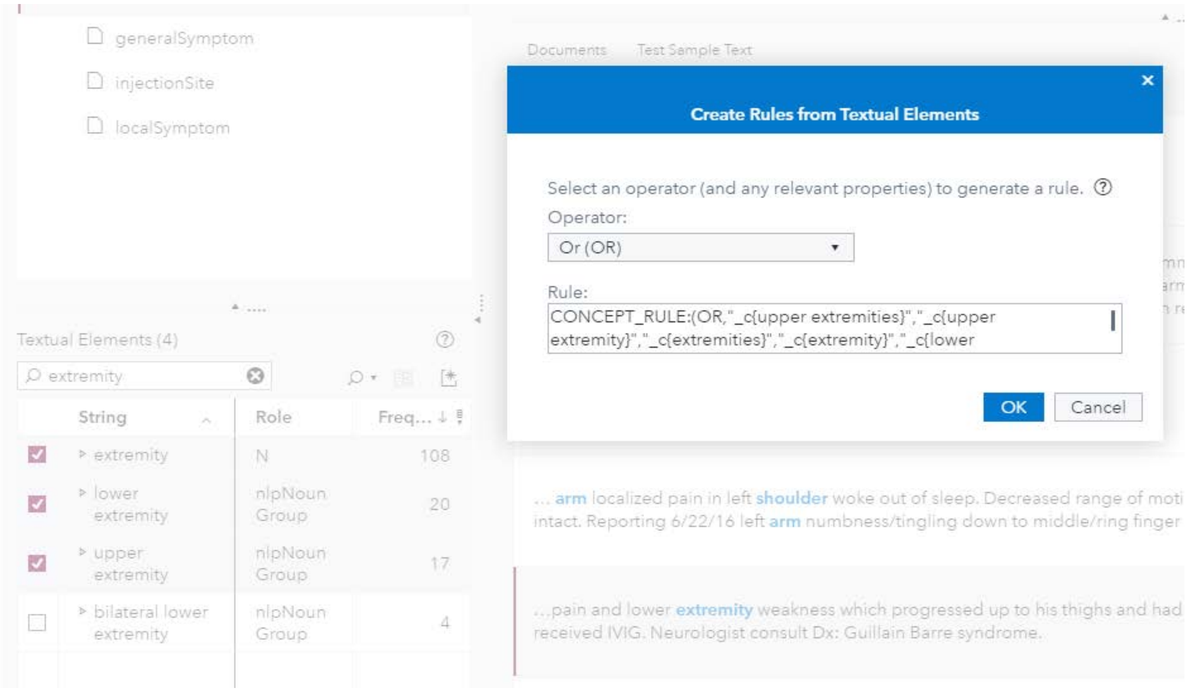
...pain and lower **extremity** weakness which progressed up to his thighs and had difficulty walking. Was seen in Emergency room, admitted to hospital and received IVIG. Neurologist consult Dx: Guillain Barre syndrome.

After you run the test using the Run Sandbox button at the top of the screen, you can see matches for your concept on all your documents, just as you can with a full test of the entire model. You can also still run a smaller test only on the sample text as before, by selecting the "Test Sample Text" icon. Note that for both full tests of all document and tests on sample text only, global rules in your model, such as REMOVE_ITEM or NO_BREAK, will only be run if they are in the concept you are testing.

Testing in the Sandbox does not use any of the "published" rules in your Edit Concept editor. So, using the Sandbox is a good way to focus in on a smaller set of rules that you want to evaluate. After you finish your evaluation, if you want to "publish" the new rules to your model, you can click on the "Move to Edit Concept" icon in the upper right-hand corner to move them or just cut and paste them between the editors.

As shown in Figure 6, the exploratory rules such as CONCEPT:right _c{_w} find many useful body parts, which are highlighted in the documents. The term "extremity", for example, was not on your original list, so you decide to investigate that term a bit more to see how it behaves in the data. In the Textual Elements window to the left, you search for the term "extremity" and find three terms that you want to add; "extremity", "lower extremity", and "upper extremity". After clicking on the checkboxes by the three parent terms, you can generate a rule that includes all three terms and their plural forms, by clicking on the icon "Create rules from selected elements".

Figure 7 Using Textual Elements and Rule Wizard



Create Rules from Textual Elements

Select an operator (and any relevant properties) to generate a rule.

Operator: Or (OR)

Rule:
CONCEPT_RULE:(OR,"_c{upper extremities}","_c{upper extremity}","_c{extremities}","_c{extremity}","_c{lower

	String	Role	Freq...
☑	extremity	N	108
☑	lower extremity	nlpNoun Group	20
☑	upper extremity	nlpNoun Group	17
□	bilateral lower extremity	nlpNoun Group	4

... arm localized pain in left shoulder woke out of sleep. Decreased range of moti intact. Reporting 6/22/16 left arm numbness/tingling down to middle/ring finger

...pain and lower extremity weakness which progressed up to his thighs and had received IVIG. Neurologist consult Dx: Guillain Barre syndrome.

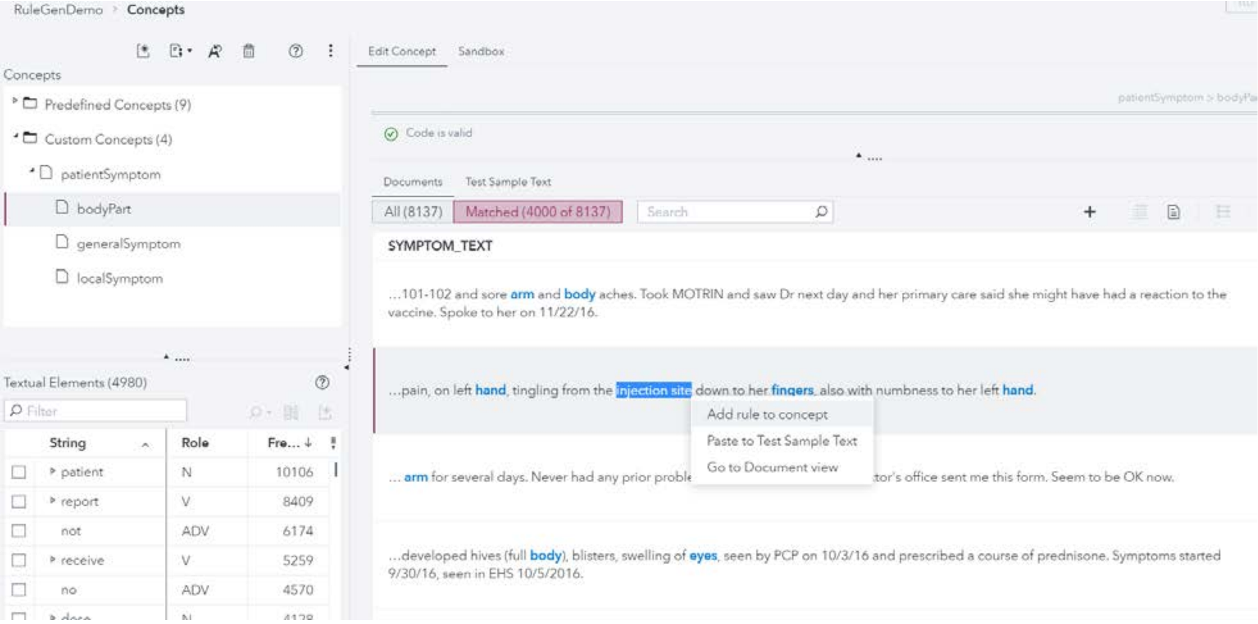
A rule wizard pops up, as shown in Figure 7, to help with creation of the rule. Note that the created rule will appear in the Sandbox, because that is the currently active workspace. You can test in the Sandbox and then move the rule once you are satisfied that you want to use it in your final model.

The Sandbox is a useful space for developing rules for a concept. Not only can you use it to test and validate new rules, but you can also use it to store less useful rules that you have commented, so that you retain a development record of your hypotheses. Be aware, however, that the two Move and Delete icons at the top will apply to your whole Sandbox for the concept, either moving all contents to the Edit Concept space or deleting all the Sandbox contents for that concept. A best practice recommendation for controlling what is moved is to select what you want to move and use the regular keyboard or right-click commands for copy and paste.

ADDING CLASSIFIER RULES FROM DOCUMENTS

As you continue to review the results for the body part concept, you notice that references to "injection site" are not matched as a body part, and you initially decide to add the term as a special rule for body type for your purpose. To easily add the rule, you can highlight the text and right-click to select from the drop-down menu: "Add rule to concept" as shown in Figure 8. Issuing this command results in a CLASSIFIER rule being appended to the end of the rules list. To investigate this special rule further, you copy and paste it from the end of your rules list to the Sandbox and look for potential variations.

Figure 8 Add Rule By Highlighting Text in Document



	String	Role	Fre... ↓
□	▸ patient	N	10106
□	▸ report	V	8409
□	not	ADV	6174
□	▸ receive	V	5259
□	no	ADV	4570

SYMPTOM_TEXT

...101-102 and sore **arm** and **body** aches. Took MOTRIN and saw Dr next day and her primary care said she might have had a reaction to the vaccine. Spoke to her on 11/22/16.

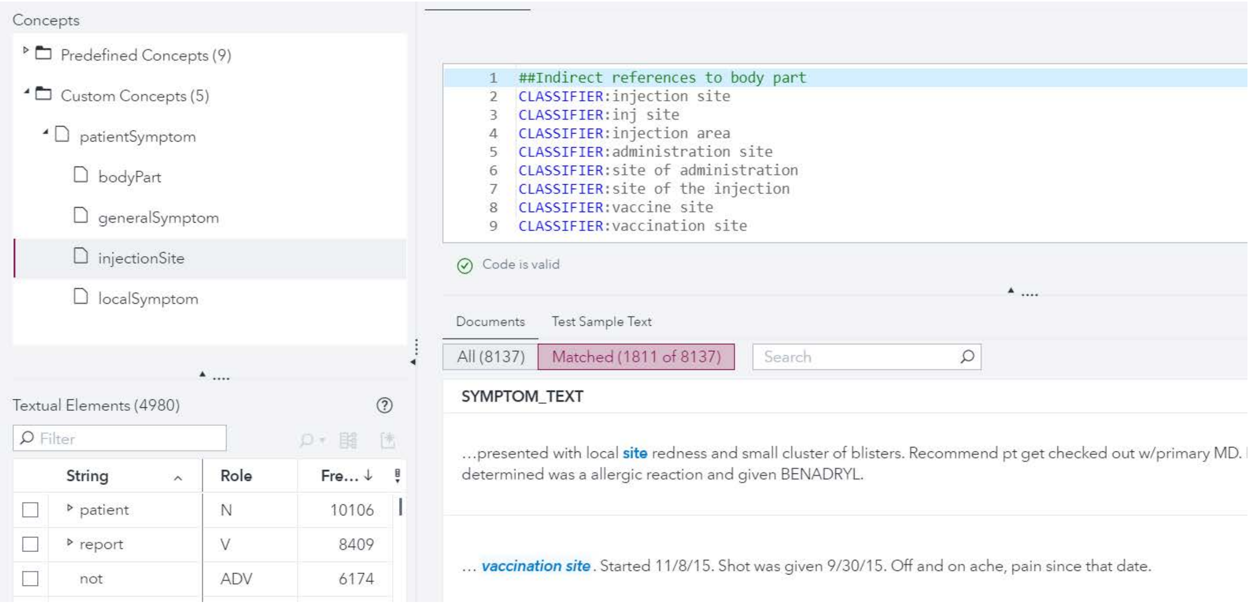
...pain, on left **hand**, tingling from the injection site down to her **fingers**, also with numbness to her left **hand**.

... **arm** for several days. Never had any prior proble tor's office sent me this form. Seem to be OK now.

...developed hives (full **body**), blisters, swelling of **eyes**, seen by PCP on 10/3/16 and prescribed a course of prednisone. Symptoms started 9/30/16, seen in EHS 10/5/2016.

To quickly find some variations, you search on the word "site" and find 1808 documents match of the 8137 in the data. Several are variations, and through investigation and use of the highlight + "add CLASSIFIER rule" feature, you add a few more rules related to the injection site. Technically, all the rules you added overlap with the one rule CLASSIFIER:site. However, for clarity, you have decided that when more information is present in the text, you will capture it. Also, because you have so many variations, you decide to separate this type of location from typical body parts in our model. Instead of moving these rules back to the Edit Concept tab, you add a new concept to your model and design documentation called injectionSite and move the rules to the new concept, as shown in Figure 9.

Figure 9 Addition of Variations After Search for "site" in New Concept



```
##Indirect references to body part
CLASSIFIER:injection site
CLASSIFIER:inj site
CLASSIFIER:injection area
CLASSIFIER:administration site
CLASSIFIER:site of administration
CLASSIFIER:site of the injection
CLASSIFIER:vaccine site
CLASSIFIER:vaccination site
```

Code is valid

SYMPTOM_TEXT

…presented with local site redness and small cluster of blisters. Recommend pt get checked out w/primary MD. I determined was a allergic reaction and given BENADRYL.

… *vaccination site*. Started 11/8/15. Shot was given 9/30/15. Off and on ache, pain since that date.

	String	Role	Fre…
□	patient	N	10106
□	report	V	8409
□	not	ADV	6174

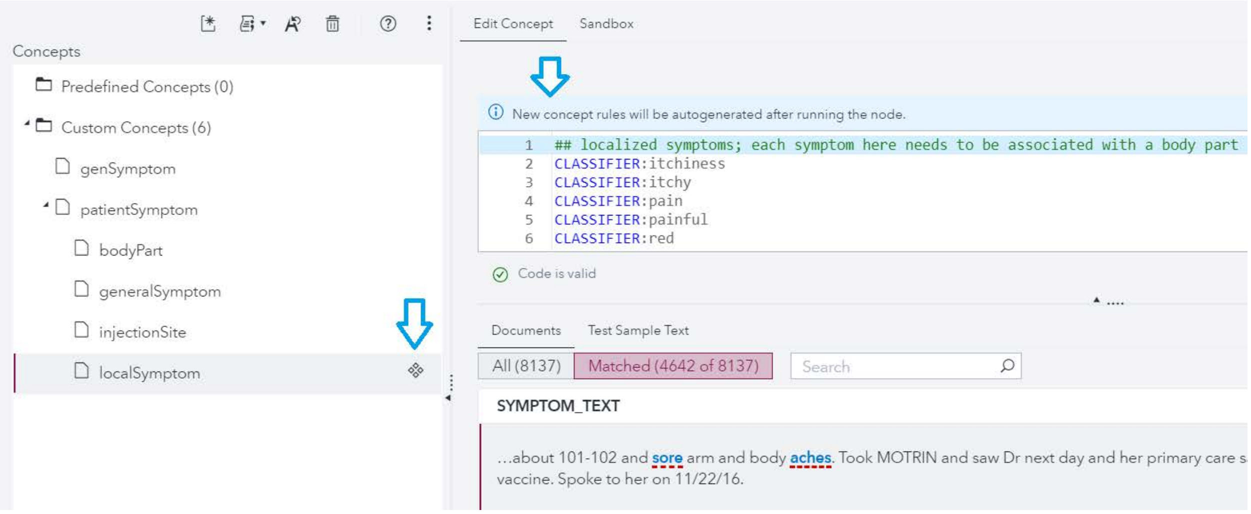
DEVELOPING SYMPTOM CONCEPTS

When working on the two symptoms lists, you approach them similarly to the body parts list, first relying on lists from your subject matter experts. For the localized symptom concept, you begin with 42 rules. Because this is a relatively short list, you want to explore to see if you might be missing some important symptoms. So, you decide to try automatic concept rule generation. This is an experimental feature using a new algorithm based on a template of n-grams. You have read the documentation and understand that you should not expect precise rules as output, but you have enough experience with LITI rule syntax to understand what to expect: a blend of CLASSIFIER, C_CONCEPT, and potentially CONCEPT_RULE rule types.

The algorithm will leverage the rules you already have in your concept and use a new table that is created by the applyConcept action called ruleMatchOut to identify patterns in the current matches. CLASSIFIER rules will be created to identify n-grams or tokens that your rules match; this is why putting a set of C_CONCEPT rules into a concept then running rule generation can help find good candidates for CLASSIFIER rules. C_CONCEPT rules will be created as n-grams (usually tri-grams) leveraging context on either side of your match and examining part-of-speech tag patterns as substitutes for exact strings. These rules can help you cast a wide net and also find useful patterns. Finally, CONCEPT_RULE rules are generated if there are predefined concepts turned on. This type of rule looks for ways to narrow your rule matches by leveraging the context of predefined concepts within the scope of a sentence; nlpNounGroup, nlpTime, nlpDate, nlpPercent, and nlpMoney are used.

Next, you mark your localSymptom concept for concept rule generation, which places an icon next to the concept in the taxonomy, and a message at the top of the Edit Concept window telling you that rules will be generated in the next run of the node – shown by the two arrows in Figure 10.

Figure 10 Concept Marked for Rule Generation



New concept rules will be autogenerated after running the node.

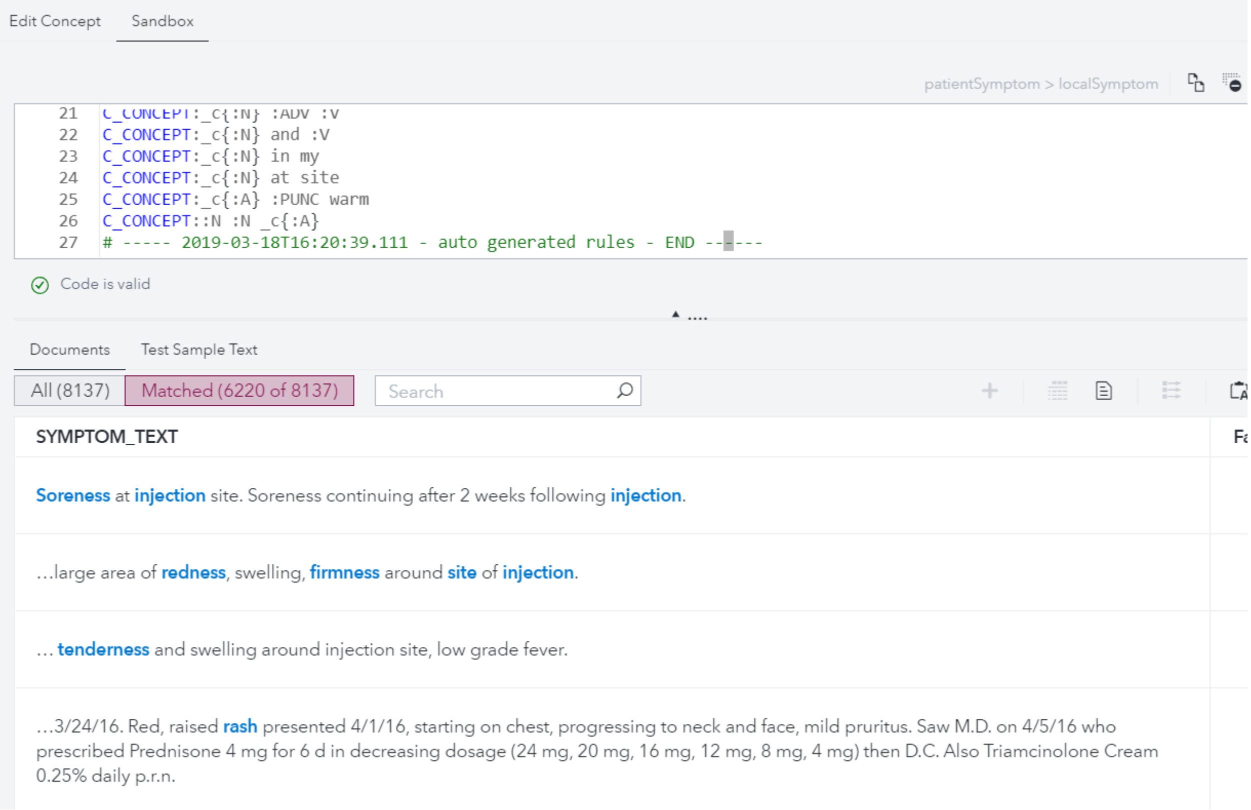
```
## localized symptoms; each symptom here needs to be associated with a body part
CLASSIFIER:itchiness
CLASSIFIER:itchy
CLASSIFIER:pain
CLASSIFIER:painful
CLASSIFIER:red
```

Code is valid

…about 101-102 and sore arm and body aches. Took MOTRIN and saw Dr next day and her primary care s vaccine. Spoke to her on 11/22/16.

You next run the node and find that the Sandbox has a blue dot next to it, indicating that you can find your new rules there. When you see the new rules, you first notice some keywords that are relevant: rash, swelling, swollen, warm, lump, hurt, ache. You note that the word "warm" was not in your original list of rules, so you add "warm" and "warmth" to your list. Then you decide to run all the rules to see if they uncover patterns you have missed; you select the "Run Sandbox" button in the upper right to run the new rules.

Figure 11 Automatically Generated Rules and Matches for localSymptom Concept



```
C_CONCEPT:_c{:N} :ADV :V
C_CONCEPT:_c{:N} and :V
C_CONCEPT:_c{:N} in my
C_CONCEPT:_c{:N} at site
C_CONCEPT:_c{:A} :PUNC warm
C_CONCEPT::N :N _c{:A}
# ----- 2019-03-18T16:20:39.111 - auto generated rules - END -----
```

Code is valid

Soreness at injection site. Soreness continuing after 2 weeks following injection.

…large area of redness, swelling, firmness around site of injection.

… tenderness and swelling around injection site, low grade fever.

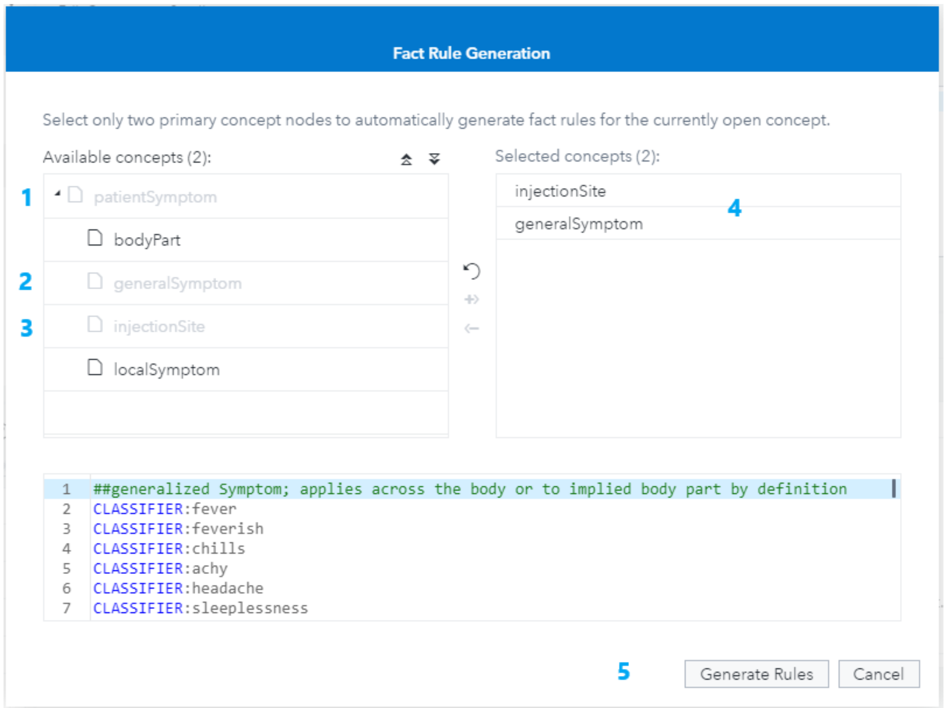
…3/24/16. Red, raised rash presented 4/1/16, starting on chest, progressing to neck and face, mild pruritus. Saw M.D. on 4/5/16 who prescribed Prednisone 4 mg for 6 d in decreasing dosage (24 mg, 20 mg, 16 mg, 12 mg, 8 mg, 4 mg) then D.C. Also Triamcinolone Cream 0.25% daily p.r.n.

The output of the run is shown in Figure 11. As expected, the highlighted items are only sometimes relevant to local symptoms and some matches only duplicates you already match with your original rules. However, you review the matched text and discover some new items to add to your list of rules, including "loss of strength", "limited range of motion", "welt@N", "lumps", "myalgia", and "circle". You reject most of the automated rules but benefit from the wide net they cast to find new items for your list.

To develop your list of generalized symptoms, you proceed in a similar fashion. You begin with about 44 rules. You use the automated rule generation to check for more items for your list, and you use the Textual Elements list to identify items that might need to be added, sorting into target nouns and noun group sections. Once you have worked on this concept for a while, you start to notice that these symptoms are often in contexts that you did not expect and might not want to extract for your reports. For example, symptoms in a negated context such as "no fever". Also, with the word "fever", sometimes you see "yellow fever" (describing a vaccine type). Finally, you notice that sometimes the symptoms described are not current symptoms, but rather part of the patient's past medical record. You handle all those issues using strategies in chapters 9.2.3 and 14.6.1 of *Text Analytics for Business Applications: Concept Rules for Information Extraction Models* by Jade, et al.

DEVELOPING FACT RULES FOR PATIENT SYMPTOM CONCEPT

When you have the lists of both body parts/injection site and localized symptoms, you want to model the relationship between the two. In other words, when the localized symptom occurs at a specified site on the body, you want to be able to tie the two items together in your output. A fact rule will do this. To start this process, you leverage the new fact rule generation feature by first selecting the concept in which you want to put the new rules, in this case the patientSymptom concept, and then selecting the menu item "Autogenerate Fact Rules", located above the taxonomy window. The "Fact Rule Generation" window pops up to help you choose the two concepts from which you want to generate the rules. Both concepts should be able to run successfully and generate matches.



Fact Rule Generation

Select only two primary concept nodes to automatically generate fact rules for the currently open concept.

Available concepts (2):

- patientSymptom
  - bodyPart
  - generalSymptom
  - injectionSite
  - localSymptom

Selected concepts (2):

- injectionSite
- generalSymptom

```
##generalized Symptom; applies across the body or to implied body part by definition
CLASSIFIER:fever
CLASSIFIER:feverish
CLASSIFIER:chills
CLASSIFIER:achy
CLASSIFIER:headache
CLASSIFIER:sleeplessness
```

Figure 12 Wizard to Select Concepts for Fact Rule Generation

First, you select the two concepts injectionSite and localSymptom, as shown in Figure 12. In the ″Fact Rule Generation″ window, you can see the concept you have selected to hold the fact rules is grayed out and cannot be selected as a source (marked by the number 1 in the figure), the two concepts selected here are marked by numbers 2 and 3 and have been moved over to the selected concepts side (marked by number 4). Finally, you can finish your selections by pressing the "Generate Rules" button (marked by number 5) to go back to your Concepts node window. When you return, you will see icons marking the concept patientSymptom to show that next time to you run the node the rules will be generated.

Again, after the node runs, the automatically generated rules show up in the sandbox for the concept you selected. In this case, you selected the patientSymptom concept, so when you go to the Sandbox for this concept, you see the following two rules:

```
PREDICATE_RULE:(concept1, concept2):(SENT, (DIST_6,
"_concept1{localSymptom}", "_concept2{injectionSite}"))
PREDICATE_RULE:(concept1, concept2):(SENT, "_concept1{localSymptom}",
"_concept2{injectionSite}")
```

Fact rule generation in VTA 8.4 will generate either one or two rules. Each rule is bounded by the SENT operator as the first operator in the rule. One rule, the base rule, will only contain this operator; the second rule shown above. This rule allows any matches of your two concepts within the scope of a single sentence in your data. If you think you want to extend this scope, you can just change SENT to another operator like ORD, DIST_n, or AND.

The other rule, shown first above, is likely to have a second operator: ORD, ORDDIST_n, or DIST_n. In this example, you see the rule generated contains a DIST_6 operator in addition to SENT. What this tells you about your data is that most of the matches between your two selected concepts are currently very close together; within 6 tokens of one another. Because the ORDDIST is not selected, you can tell that when checked the ordering of the two concepts varied; there was no clear pattern of one concept coming before the other.

In this case, the first rule looks the most promising, so you comment the base rule and run the Sandbox to see the matches of the second rule. After you are satisfied that this rule is a good start, you repeat the steps above to run for a new fact between the bodyPart and localSymptom concepts; these are the two concepts you select in the wizard. When the run is complete, you again have two rules with your two new concepts:

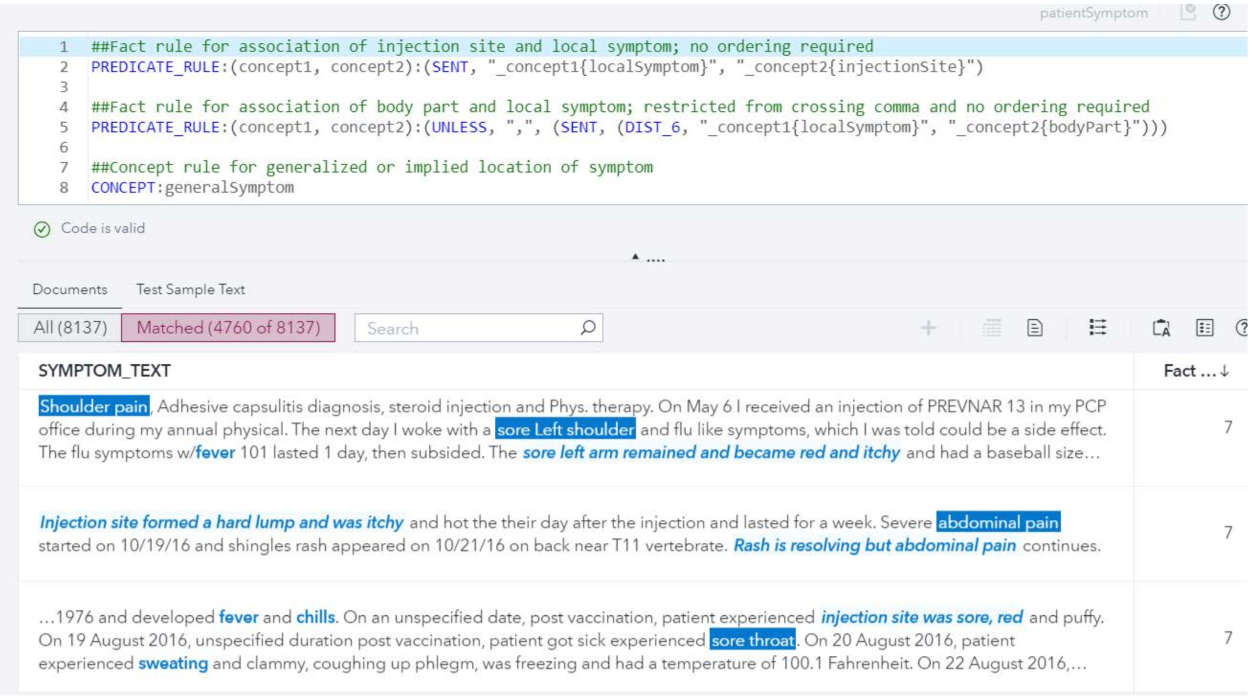
```
PREDICATE_RULE:(concept1, concept2):(SENT, (DIST_6,
"_concept1{localSymptom}", "_concept2{bodyPart}"))
PREDICATE_RULE:(concept1, concept2):(SENT, "_concept1{localSymptom}",
"_concept2{bodyPart}")
```

These rules look just like the previous two, which is reassuring, since the patterns you expect from your investigations into the injectionSite concept should parallel the bodyPart concept. You proceed to run your test of the first rule, as before, the results look promising, but seem to range a bit too far at times. You decide to work on some issues like preventing the match across commas, so you try modifying the rule by adding an UNLESS operator:

```
PREDICATE_RULE:(concept1, concept2):(UNLESS, ",", (SENT, (DIST_6,
"_concept1{localSymptom}", "_concept2{bodyPart}")))
```

After you make these changes to the rule, and run the tests in the sandbox, you move the rule over to the Edit Concept editor window with your first fact rule and the concept rule that points to the generalSymptom concept, as shown in Figure 13.

Figure 13 patientSymptom Concept Rules and Example Matches



```
##Fact rule for association of injection site and local symptom; no ordering required
PREDICATE_RULE:(concept1, concept2):(SENT, "_concept1{localSymptom}", "_concept2{injectionSite}")

##Fact rule for association of body part and local symptom; restricted from crossing comma and no ordering required
PREDICATE_RULE:(concept1, concept2):(UNLESS, ",", (SENT, (DIST_6, "_concept1{localSymptom}", "_concept2{bodyPart}")))

##Concept rule for generalized or implied location of symptom
CONCEPT:generalSymptom
```

Code is valid

SYMPTOM_TEXT	Fact ...↓
Shoulder pain, Adhesive capsulitis diagnosis, steroid injection and Phys. therapy. On May 6 I received an injection of PREVNAR 13 in my PCP office during my annual physical. The next day I woke with a sore Left shoulder and flu like symptoms, which I was told could be a side effect. The flu symptoms w/**fever** 101 lasted 1 day, then subsided. The ***sore left arm remained and became red and itchy*** and had a baseball size…	7
Injection site formed a hard lump and was itchy and hot the their day after the injection and lasted for a week. Severe abdominal pain started on 10/19/16 and shingles rash appeared on 10/21/16 on back near T11 vertebrate. ***Rash is resolving but abdominal pain*** continues.	7
…1976 and developed **fever** and **chills**. On an unspecified date, post vaccination, patient experienced ***injection site was sore, red*** and puffy. On 19 August 2016, unspecified duration post vaccination, patient got sick experienced sore throat. On 20 August 2016, patient experienced **sweating** and clammy, coughing up phlegm, was freezing and had a temperature of 100.1 Fahrenheit. On 22 August 2016,…	7

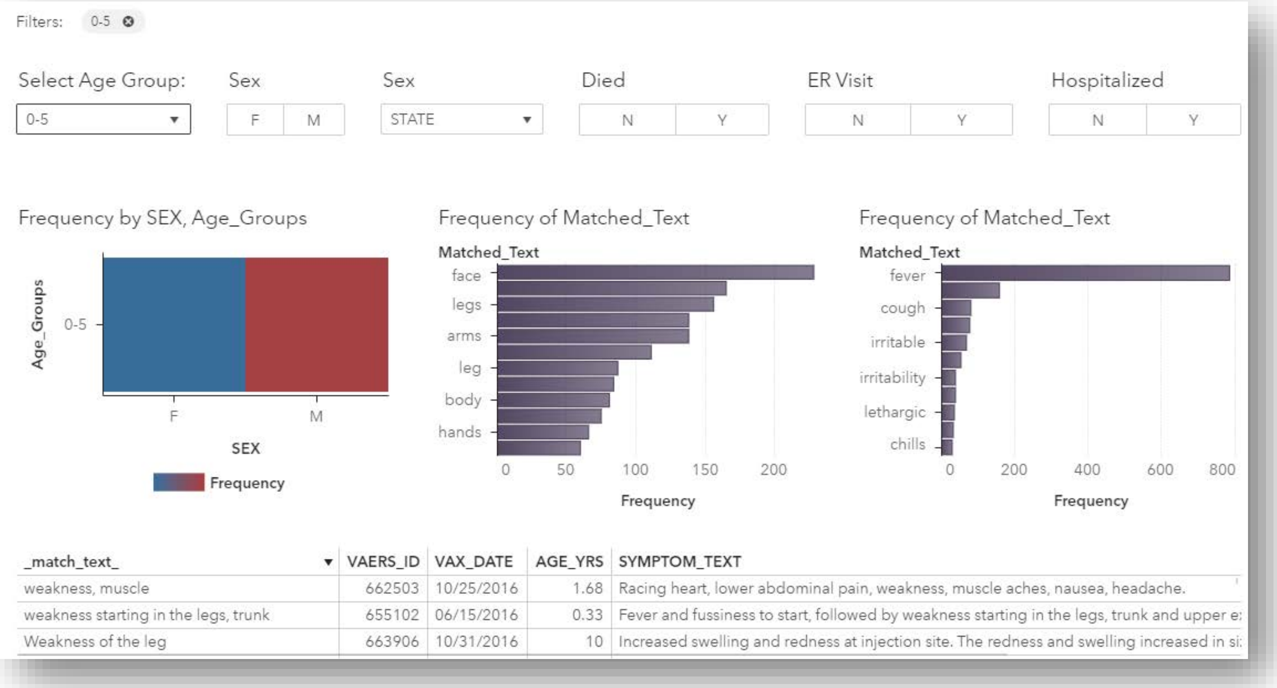
You will have some more work to do to manage the lists and coordinated items with "and" for both body parts and localized symptoms, but your current strategy gives you fine-grained control and a good starting point from which to refine your model.

CONCLUSION

This paper has demonstrated one example business case for information extraction and shown a few new ways to generate rules for your models. The process for leveraging the new Sandbox development space was illustrated, as well as three ways to generate rules for your models: a CLASSIFIER rule from highlighting, PREDICATE_RULE rules from two concepts, and other types of concept rules from a set of starter rules.

After fully building your model, the new information about symptoms for patients can be used in reports to analyze different subgroups and explore the side effects that are most prevalent in each. Any single patient can still be investigated as well as any subgroup as demonstrated in Figure 14, which shows the age group of 0-5-year-olds. The analyst can drill down by symptom, body part, gender, and see examples of the symptom text in context all at once.

Figure 14 SAS Visual Analytics Report Focusing on Children Ages 0-5



Select Age Group: 0-5

Sex: STATE

_match_text_	VAERS_ID	VAX_DATE	AGE_YRS	SYMPTOM_TEXT
weakness, muscle	662503	10/25/2016	1.68	Racing heart, lower abdominal pain, weakness, muscle aches, nausea, headache.
weakness starting in the legs, trunk	655102	06/15/2016	0.33	Fever and fussiness to start, followed by weakness starting in the legs, trunk and upper e
Weakness of the leg	663906	10/31/2016	10	Increased swelling and redness at injection site. The redness and swelling increased in si

REFERENCES

"Vaccine Adverse Event Reporting System (VAERS)". VAERS is co-sponsored by the Centers for Disease Control and Prevention (CDC), and the Food and Drug Administration (FDA), agencies of the U.S. Department of Health and Human Services (HHS). 2016. Available at https://vaers.hhs.gov/about.html.

Jade, Teresa, Biljana Belamaric Wilsey, Michael Wallis. Text Analytics for Business Applications: Concept Rules for Information Extraction Models. SAS Press, 2019.

RECOMMENDED READING

- *Text Analytics for Business Applications: Concept Rules for Information Extraction Models*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Teresa Jade
Teresa.Jade@sas.com
https://www.linkedin.com/in/teresajade

Xu Yang
Xu.Yang@sas.com
https://www.linkedin.com/in/xu-yang-8249666

Aaron Arthur
Aaron.Arthur@sas.com
https://www.linkedin.com/in/adarthur