DROP_REPORTS data set, 1414 files (doc, pdf, txt) for text analytics shown below, includes commentaries about depression/anxiety, drugs/side effects.

The comments were posted in an online forum, and the names of the medications that are referenced in the documents were altered.
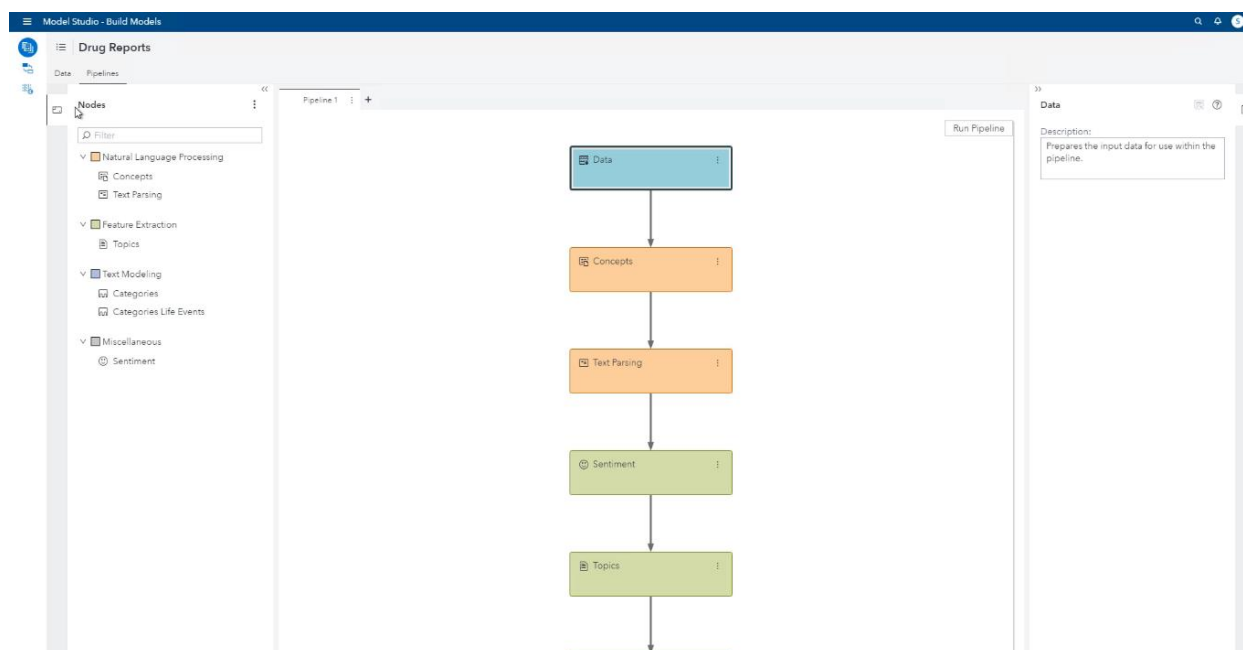
- CATEGORY Taxonomy: Topics in the document collection include the following concerns: depression and anxiety, weight gain, and sleep issues.
- CONCEPT Taxonomy: Some of the documents include the following key details: medication names, dosage levels, and side effects experienced. Four custom concepts are defined in the Concepts node.
    - DOSAGE with examples like 20 mg, 60 MG, 50 Mg, etc.
    - MEDICATION with examples like Abidal, Escalan, Cenerol, etc.
    - PRESCRIPTION with examples like Abidal 20 mg, 60 MG of Escalan, 50 Mg Cenerol, etc.
    - SIDEEFFECTS with examples like Headache, Blurred vision, Nausea, etc.

Goal is to automatically extract information from the document collection. For example, we'd like to obtain reports mentioning specific medications, so we have a Medication concept. When we score with respect to that concept, it's going to pull out all the medication names in the report. We will then combine that with dosage to get something called a Prescription. These concept rules will allow us to automatically parse the collection and create a table that has prescriptions in it. So we can immediately identify documents that mention a specific prescription, which would be a drug and a dosage.

Nodes include CONCEPTS, TEXT PARSING, SENTIMENT, TOPICS and CATEGORY
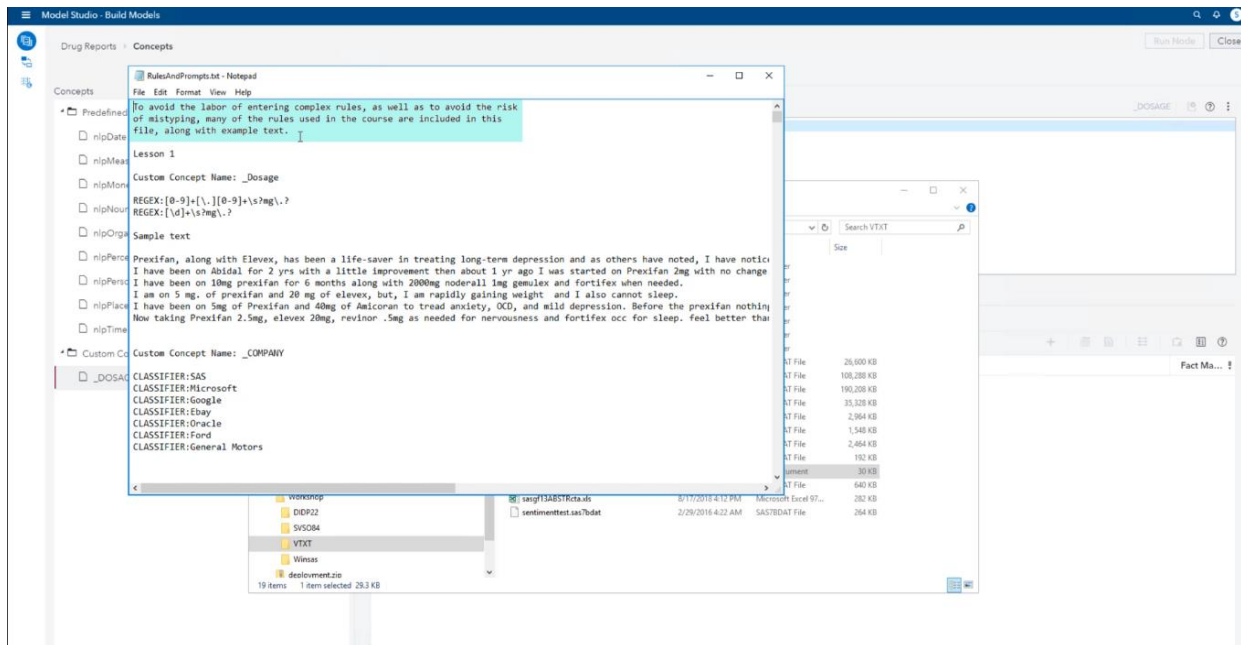

1-) Working with CONCEPTS node

Predefined concepts include many, e.g. nlpMeasure consists of measurement-like expressions; 1 year, 40 mg

There are 784 matches out of 1414 documents

RulesAnsPrompts.txt file includes conditions, e.g. to open a new concept, _DOSAGE is created with REGEX LITI Language

REGEX accepts any integer or float number with or without space, and must have "mg" and with or without "."

Function matches 300 out of 1414 documents
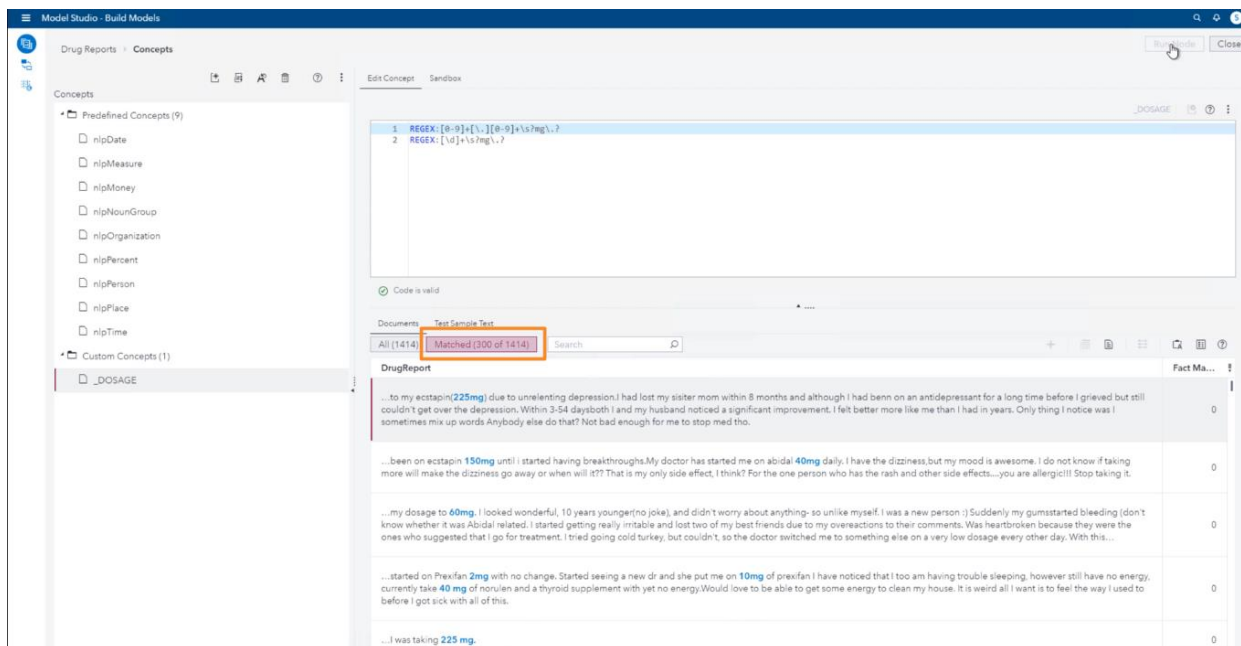
MEDICATION custom concept is created with predefined classifiers, 704 matches have been identified

2-) Working with TEXT PARSING Node

Term "depression" has 492 documents. The drug abidal is directly associated with depression based on co-occurrence statistics. A total of 292 documents contain the drug name abidal, and 142 of those abidal documents also contain the word depression. The line that connects depression and abidal has a thickness that is relative to the information gain (IG). The largest information gain is associated with anxiety. Of all the medications in the MEDICATION concept, abidal has the largest information gain as depicted by the term map.

Term Map

The medication **imitap** has the highest similarity scores (0.396) relative to depression. Some medications have negative associations.

Confusion has a similarity score of 0.224, and tiredness has a negative similarity score of -0.177. This would imply that tiredness is not a side effect of abidal.

There's a relatively high frequency of MEDICATION, 1,561 matches, SIDE_EFFECTS, 703 matches, DOSAGE, 511 matches, and PRESCRIPTION, 55 matches.

Often interested in average number of matches per document. Look at the Average Number of Matches Per Document chart. In those documents that contain prescriptions, there is approximately one prescription per document.
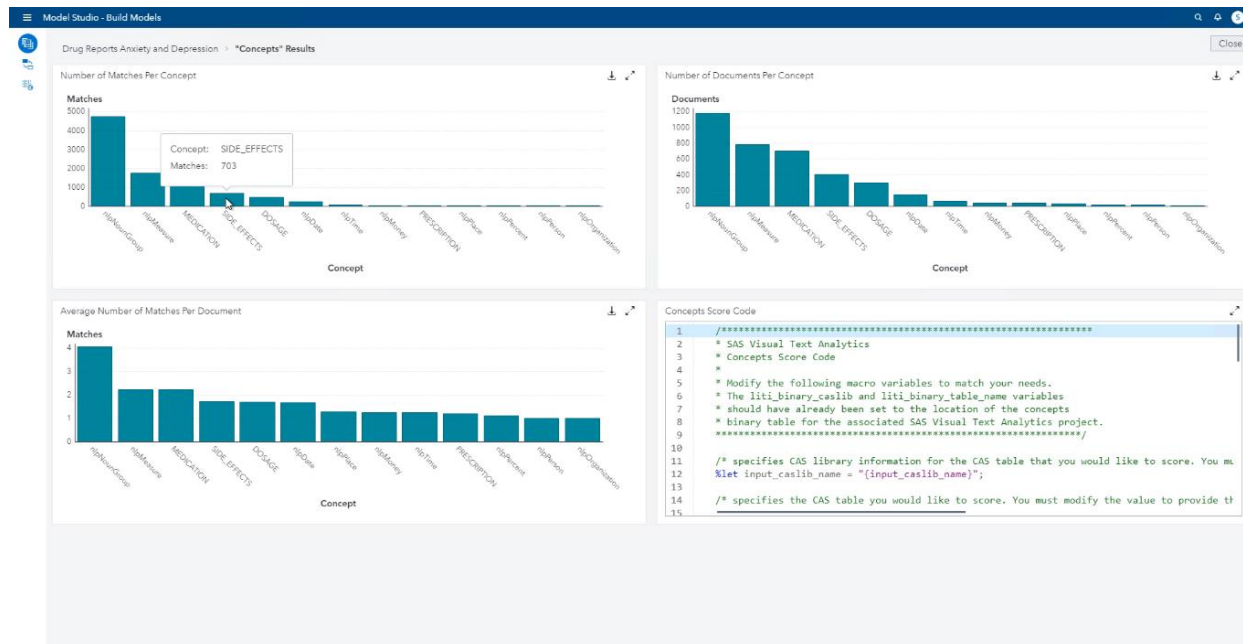
3-) Working with TOPICS node.

There are system-generated topics and the associated terms. e.g **+help, +depression, anxiety, really, +drug** topic has only 153 out of 1414 documents and 52 out of 1528 terms matches, whereas the term depression alone was found in 492 documents.
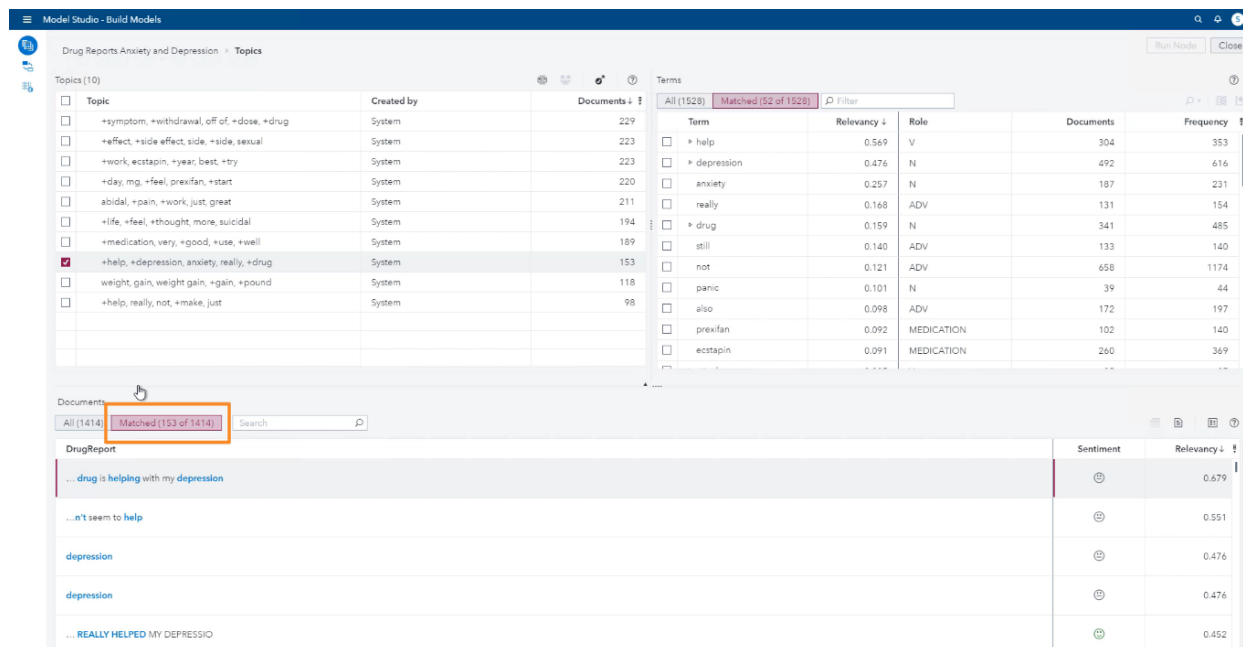
To capture more documents associated with depression and anxiety, in the second run, some setting changes were applied:

- o Clear the **Automatically determine the number of topics** box, change the **Maximum topics** value to **20**.
- o Increase the Term Density from **1** to **3**. (This increases the term relevancy threshold value.) This results in fewer significant terms associated with the topic.
- o Increase the Document Density from **1** to **2**. (This increases the document relevancy threshold value.) This results in fewer significant matched documents associated with the topic.

Increasing the maximum number of topics resulted in 20 topics. **+depression, anxiety, ecstapin, +suffer, panic** topic is selected. The number of matched documents associated with this revised topic decreased to 54. The term relevancy threshold value increased when 21 terms were associated with this topic. However increasing the number of maximum topics did not help identify more depression- and anxiety-related documents.

With the same set of topic selected, **depression** in the Terms pane has 266 matching documents.

4-) Working with CATEGORIES

One of our goals is to categorize documents with respect to *depression*. Linguistic rules and Matched documents are shown below, there are 326/1414 of them and sentiment breakdown vary.

The linguistic rules lead to 135 false positives in identifying the depression topic, 191 true positives, and 75 false negatives. The precision is around 59%. Recall is around 72%. And the F measure is around 65%. This provides the foundation for exploration.