# Week 5 Assignment – Rapid Miner

By: Zach Adair

Regis University

MSDS 650 – Data Analytics

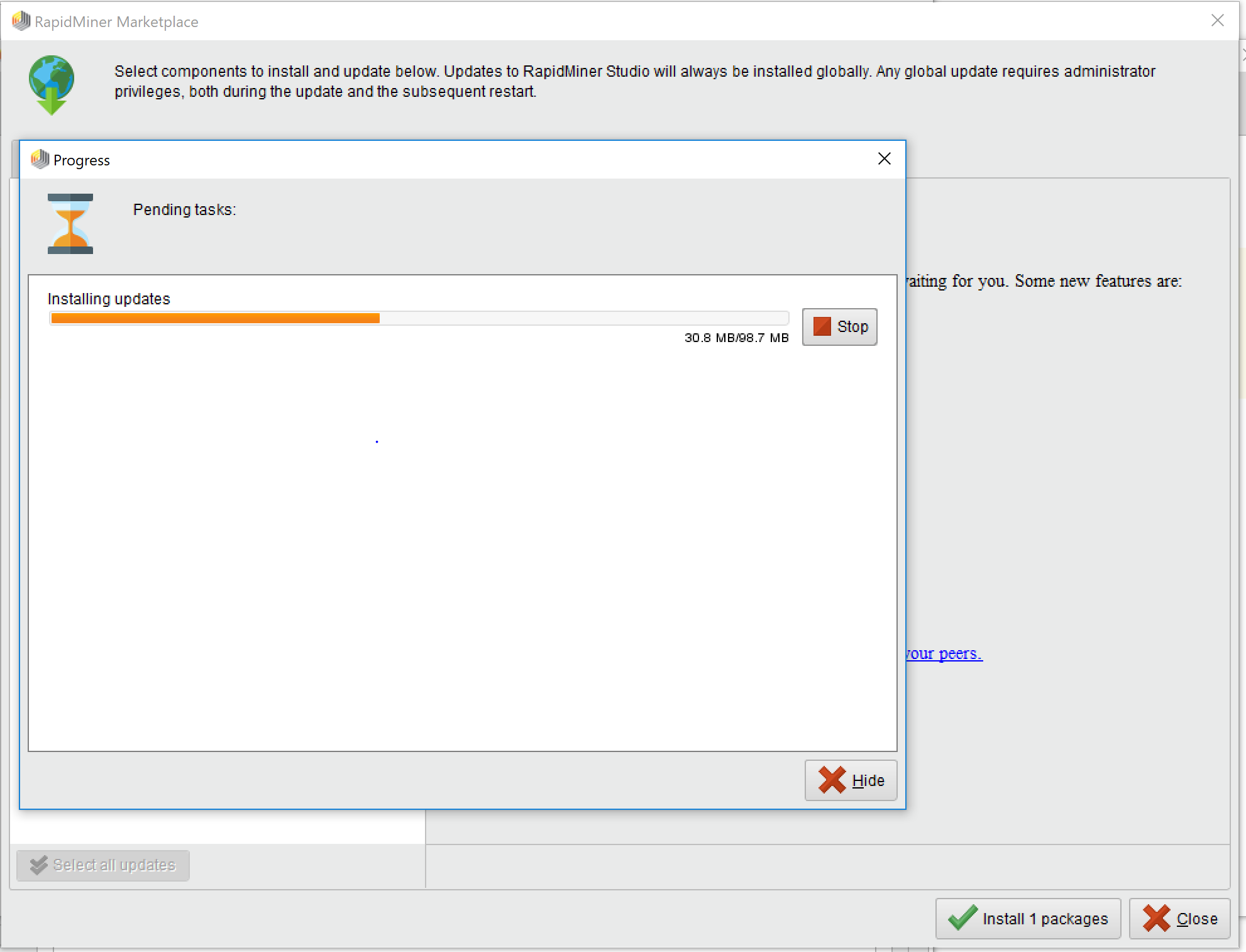
## Introduction

This week we are working with Rapid Miner, an analytics tool which has a lot of different capabilities for many different types of analysis. Rapid Miner is a tool I’m very unfamiliar with and because of that I will be working through the tutorials and exercises to make sure I am comfortable with the tool before moving onto the actual assignment. Beforehand I will be walking through the exercises to make sure I can do those properly.

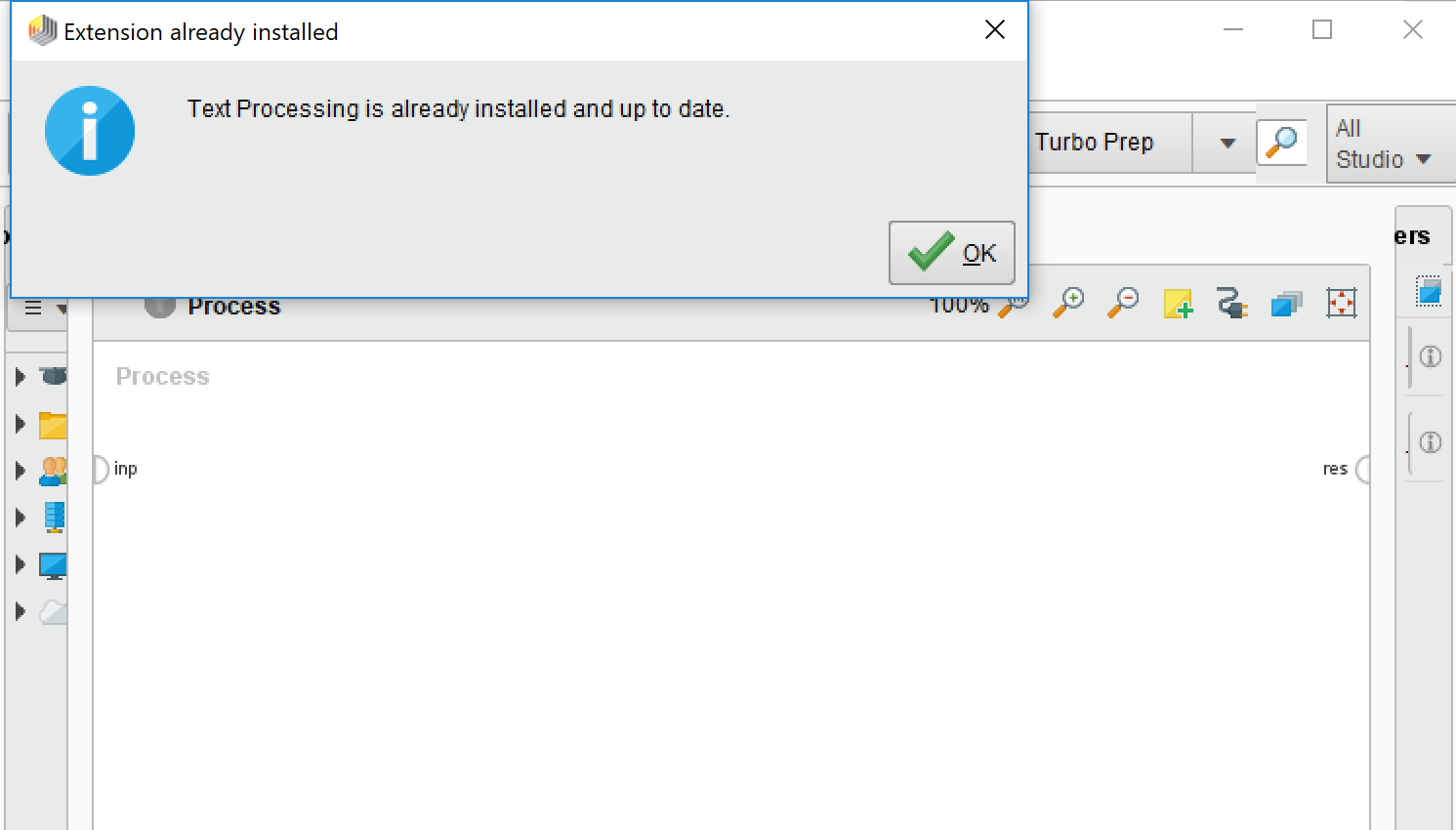
## Exercise 1:

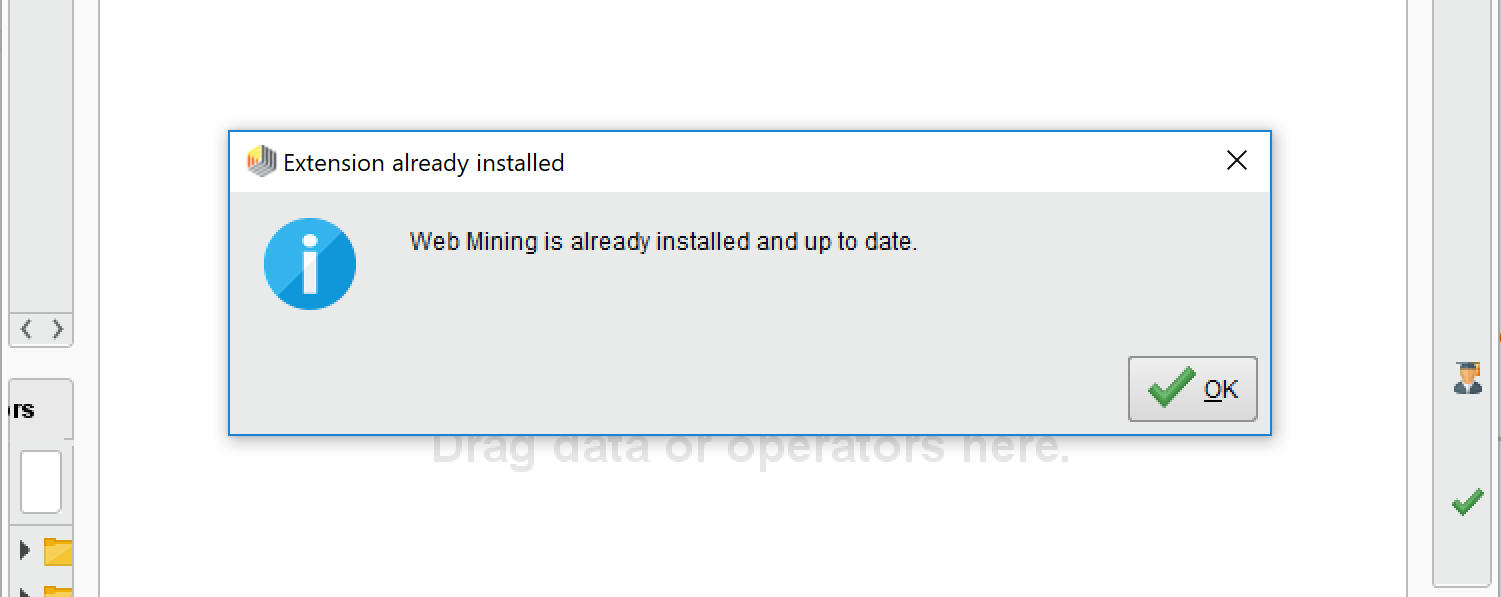
**Step 1: Install the Text Processing and Text Analysis by Aylien Extensions.**

I didn’t need to download Rapid Miner since I had done that the first week of the course but I did need to download the Text Processing and Text Analysis extension for the application, but I did need to update my version.



Next I went to the link given to upload the Text extensions.

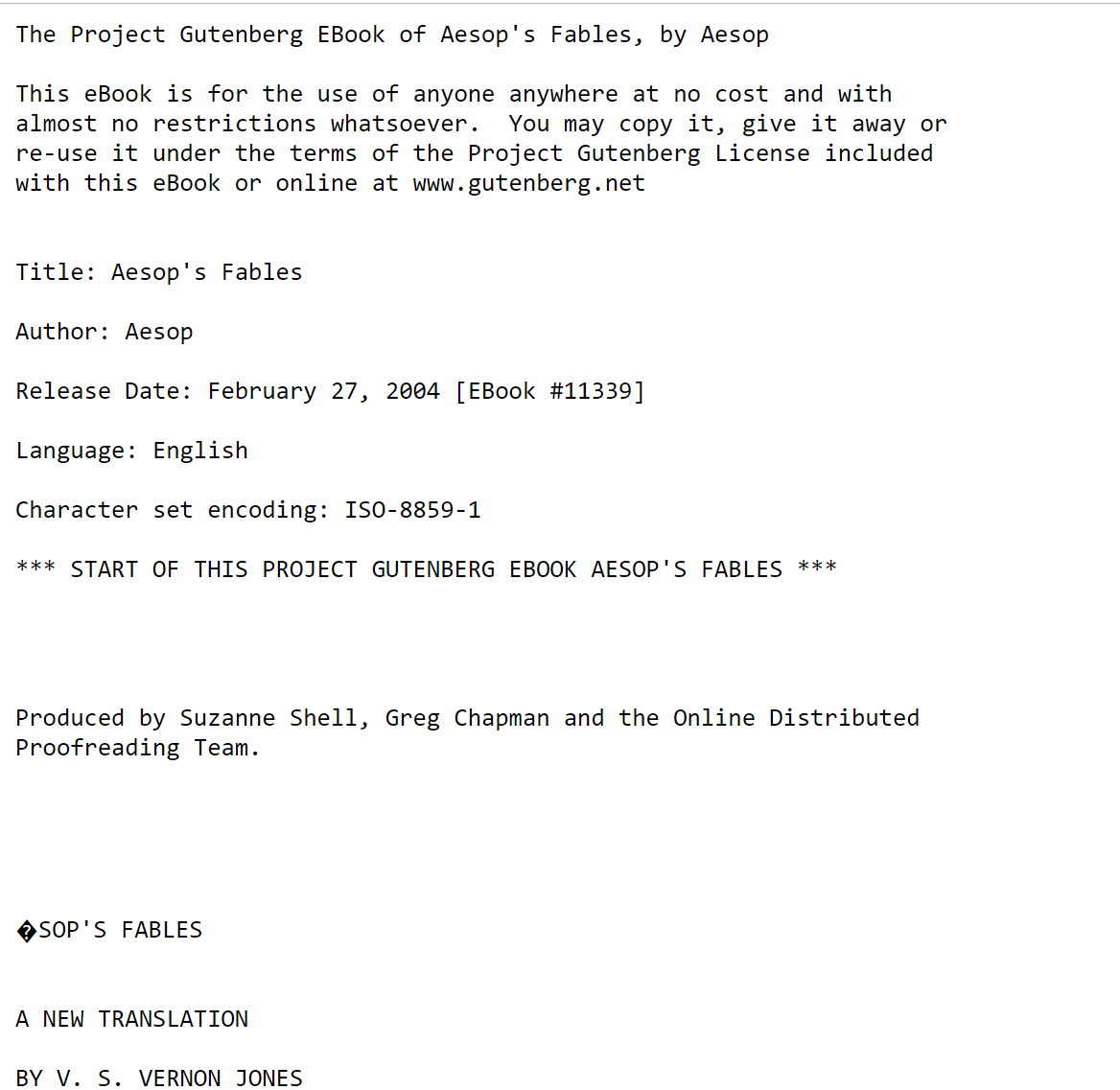




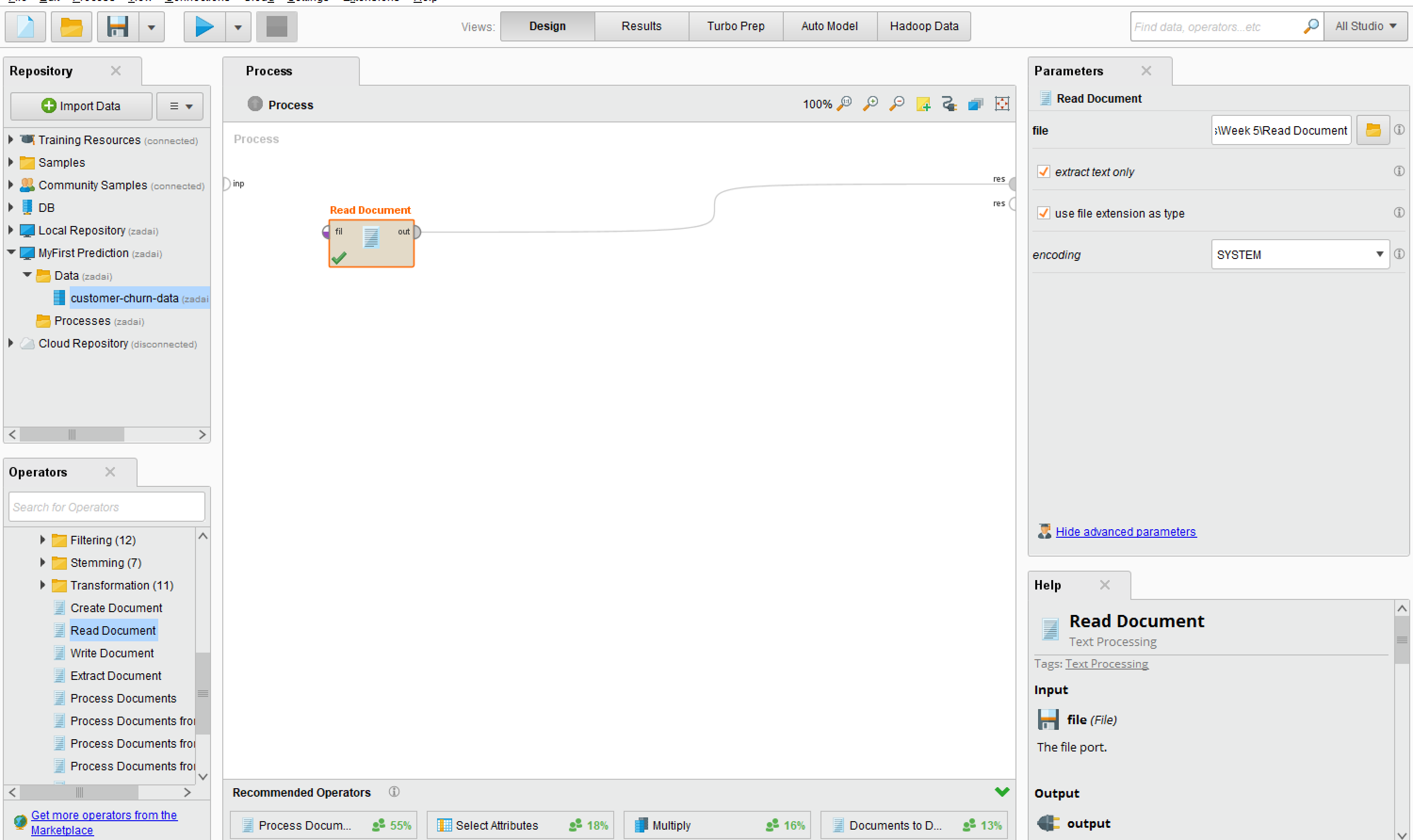
Lucky for me I had already downloaded them and I just forgot to get the screen shots of it.

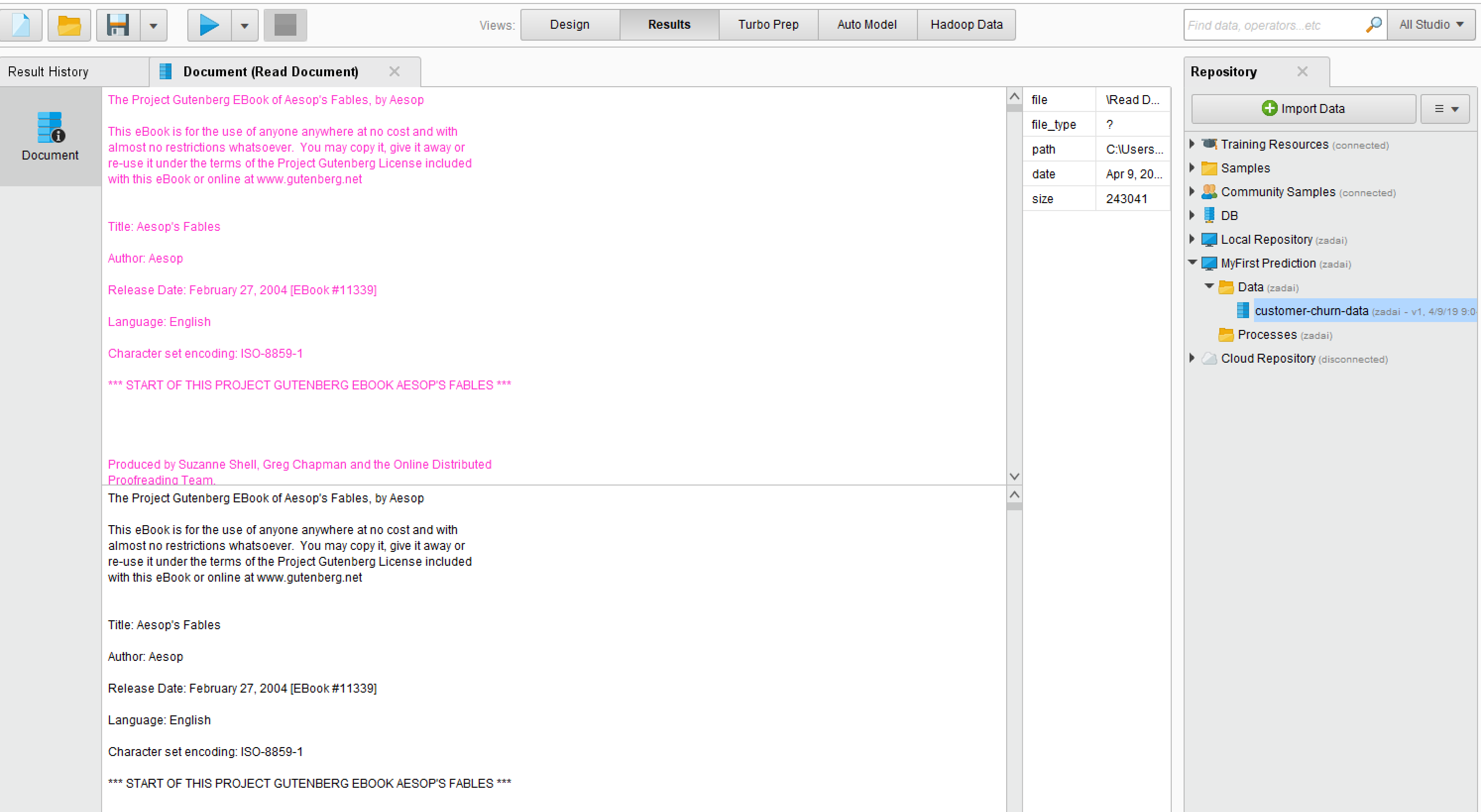
**Step 2: You can download a text of interest from gutenburg.org or any other site**

The text that I decided on downloading was Aesop’s Fables.

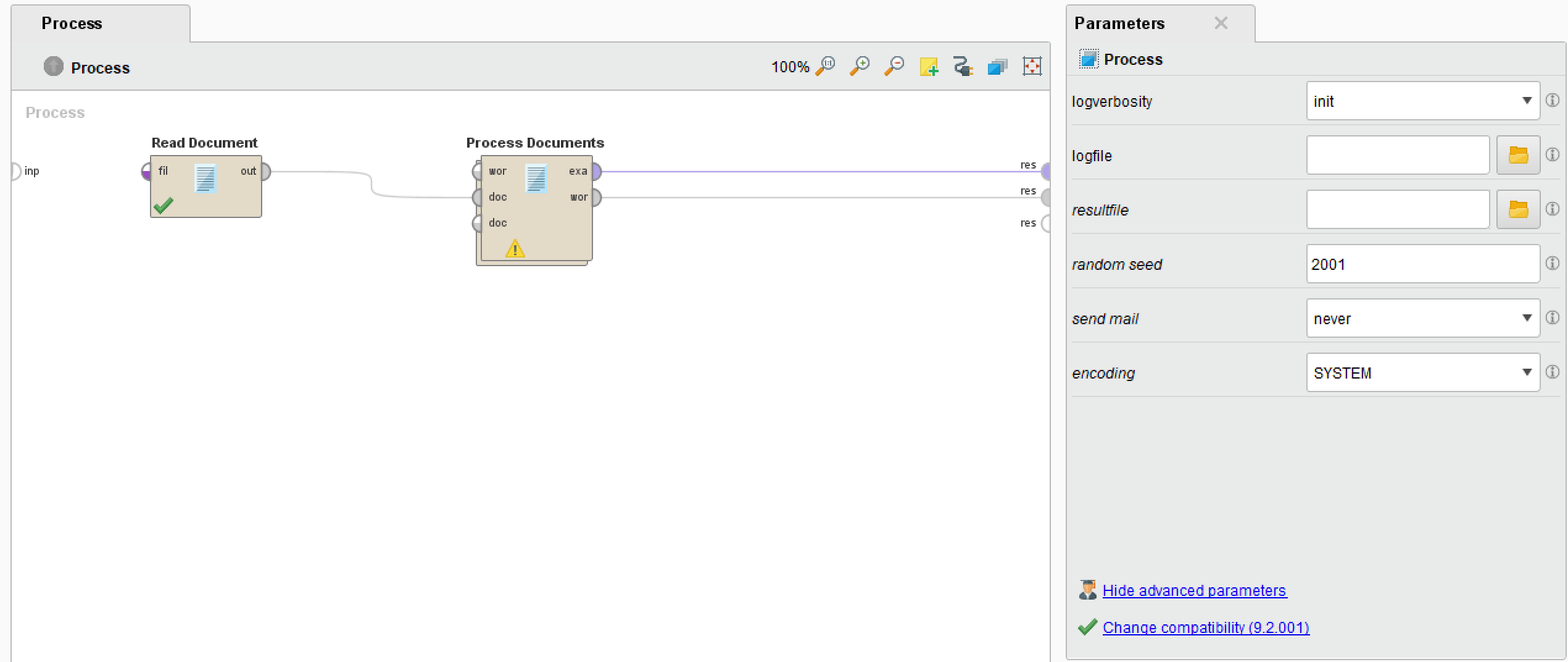


Use the file extender to read the text file that I downloaded by downloading it into Rapid Miner:

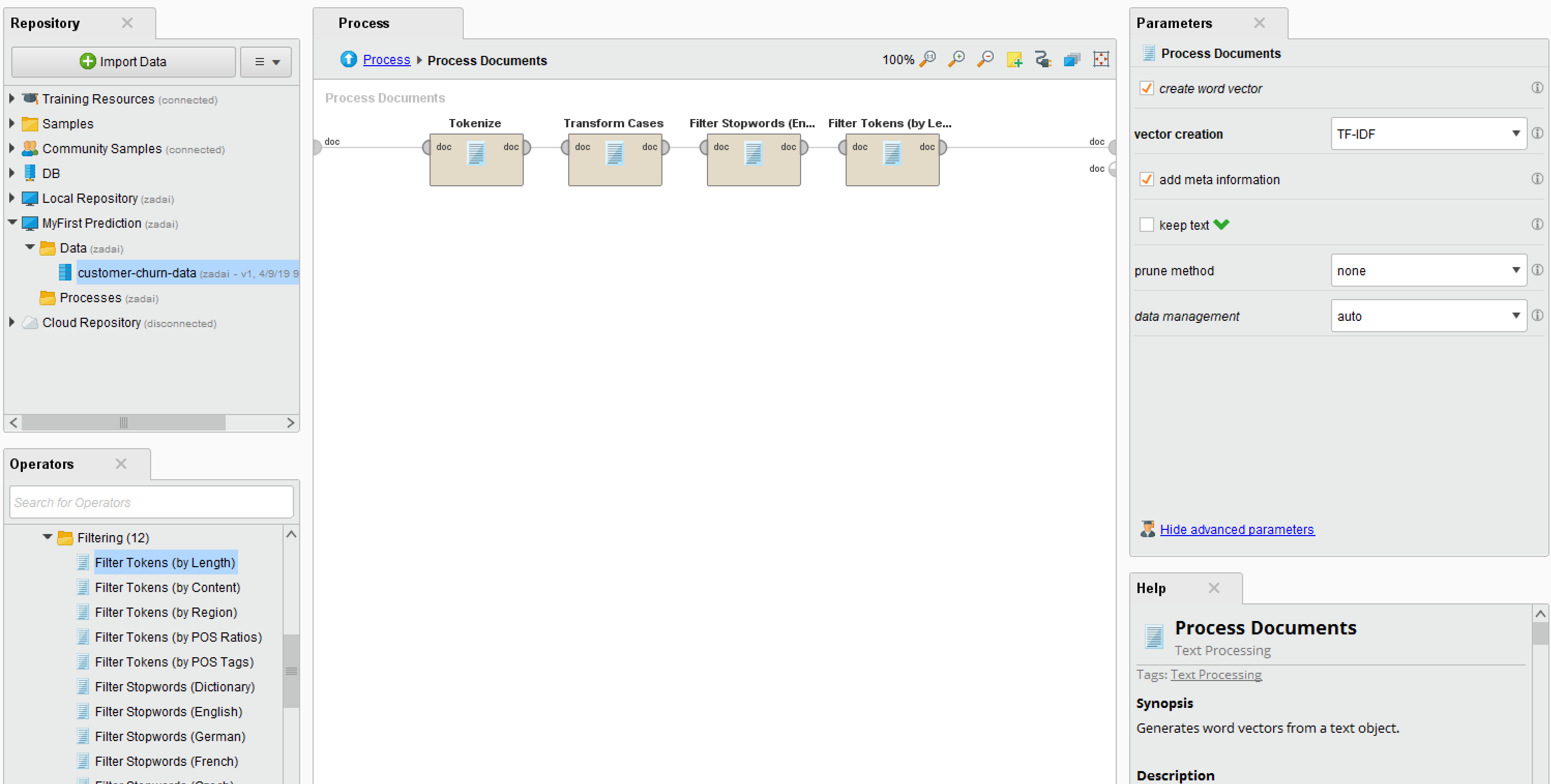




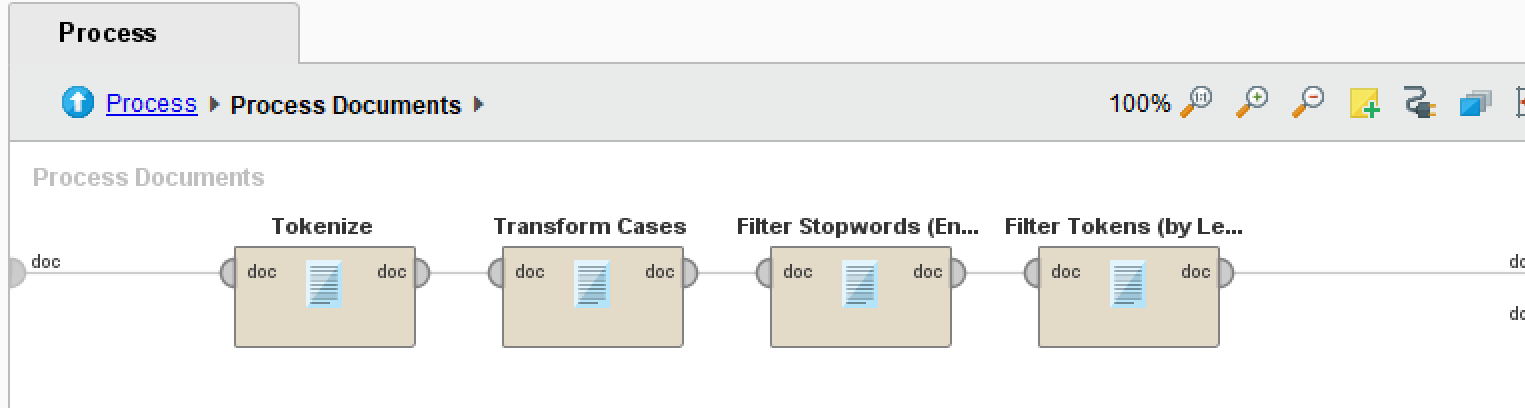
**Step 3: Bring the Process Documents operator into Main Process window and place it on the line between the Read Document and the res node**



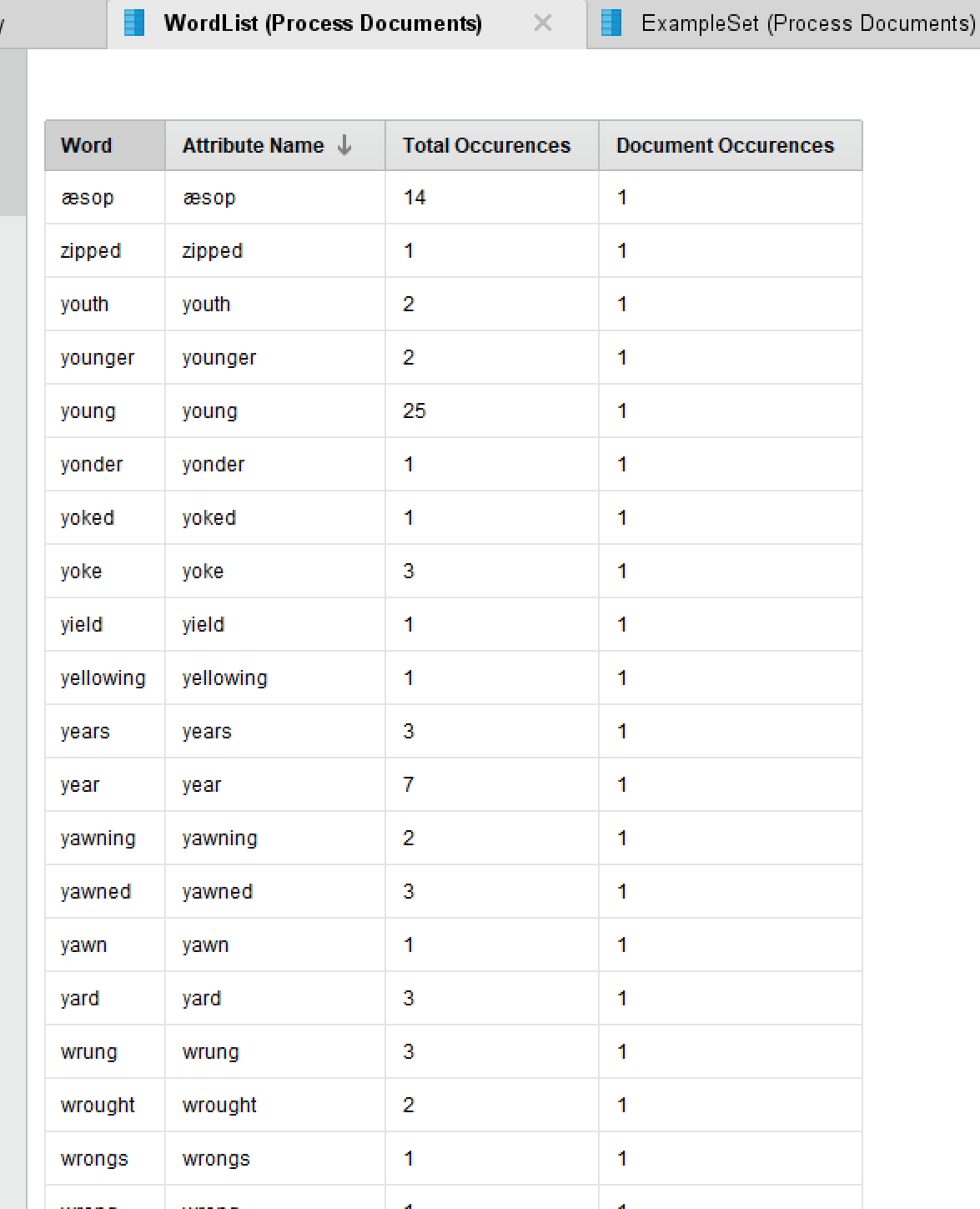
**Step 4: Double click on the Process Documents operator. This will take you inside the Process Documents operator and will allow you to add modifiers.**



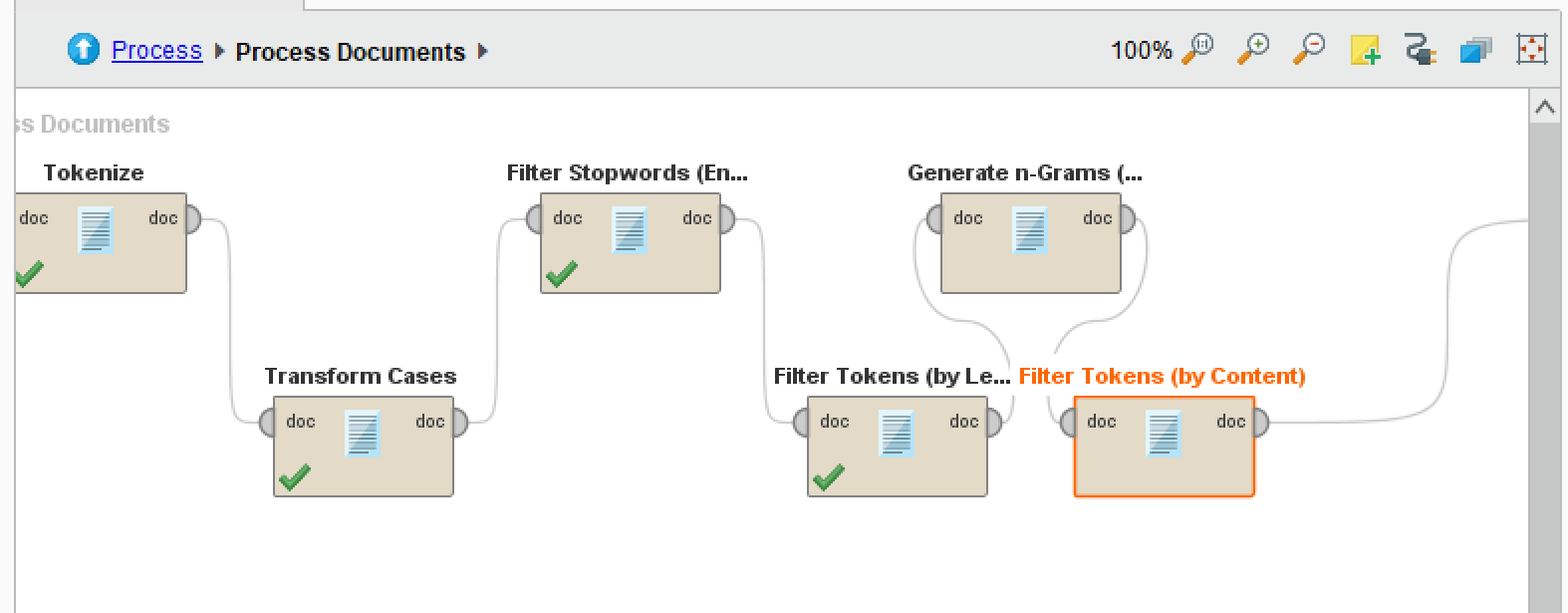
**Step 5: Return to the Main Process Window by hitting the upward pointing blue arrow or clicking the ‘Process’ blue link**



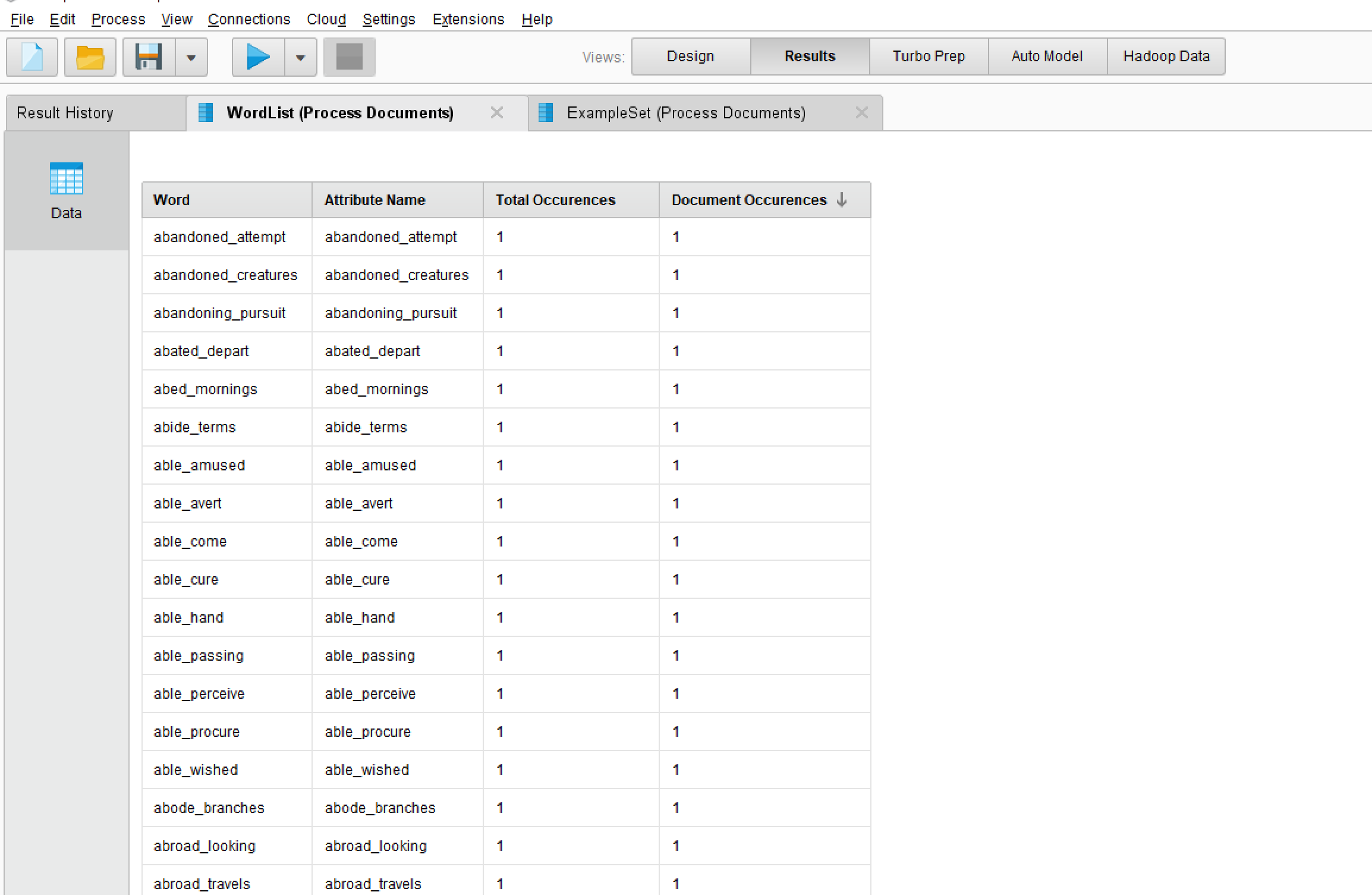
Then press “Play” button to run the process.



**Step 6: N-Grams** N-grams are combinations of n number of tokens (words) in a row. Add the Generate n-Grams (Terms) operator to your sub process.



Hit play and look at your results

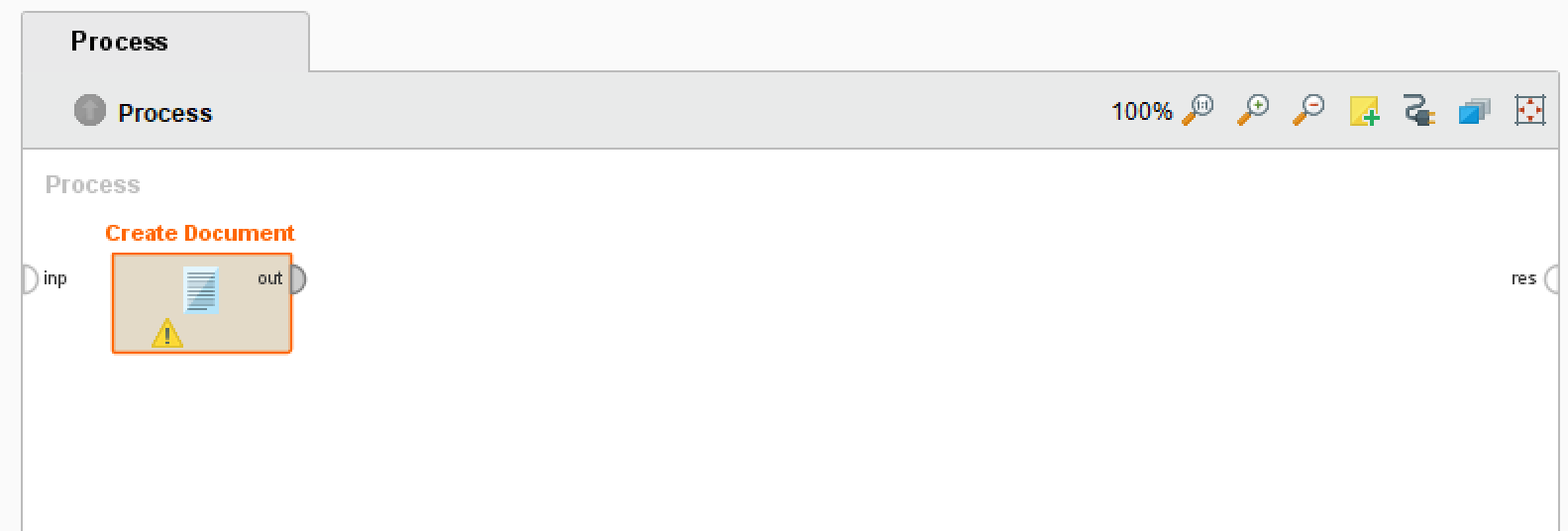


## Exercise 2: Sentiment Analysis

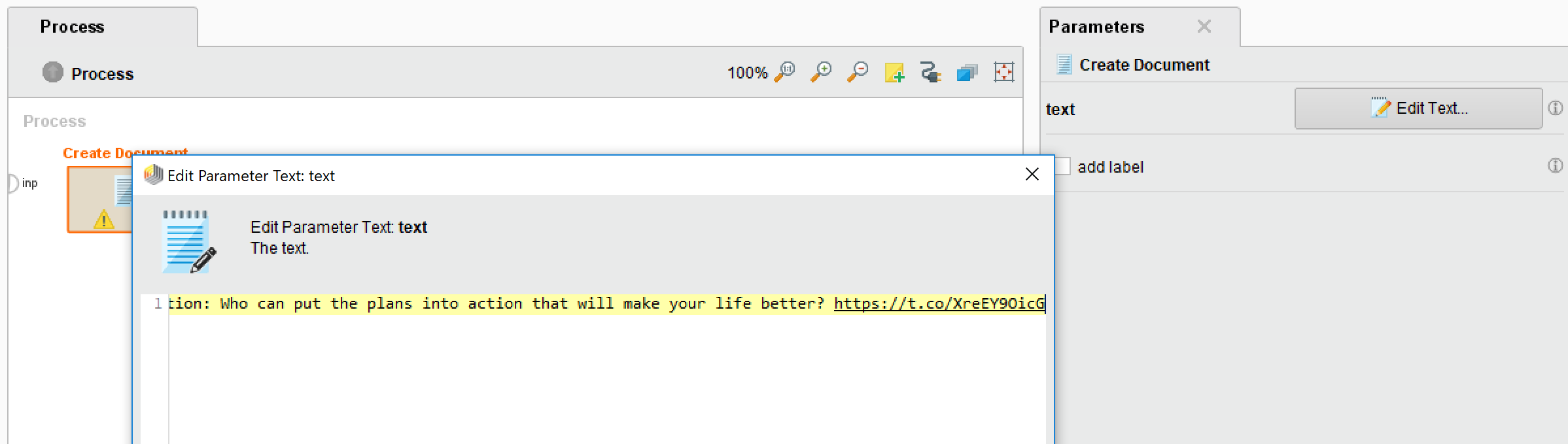
**Introduction:** We’re going to evaluate the overall sentiments of Hillary Clinton and Donald Trump’s tweets , to see how negative/positive their tweets are in general. Use the attached file trumptastic\_hildog\_tweets.xlsx. We’re going to start by testing just one tweet, so open the .csv and copy one of the tweets.

**Step 1: Setting up the test input text.**

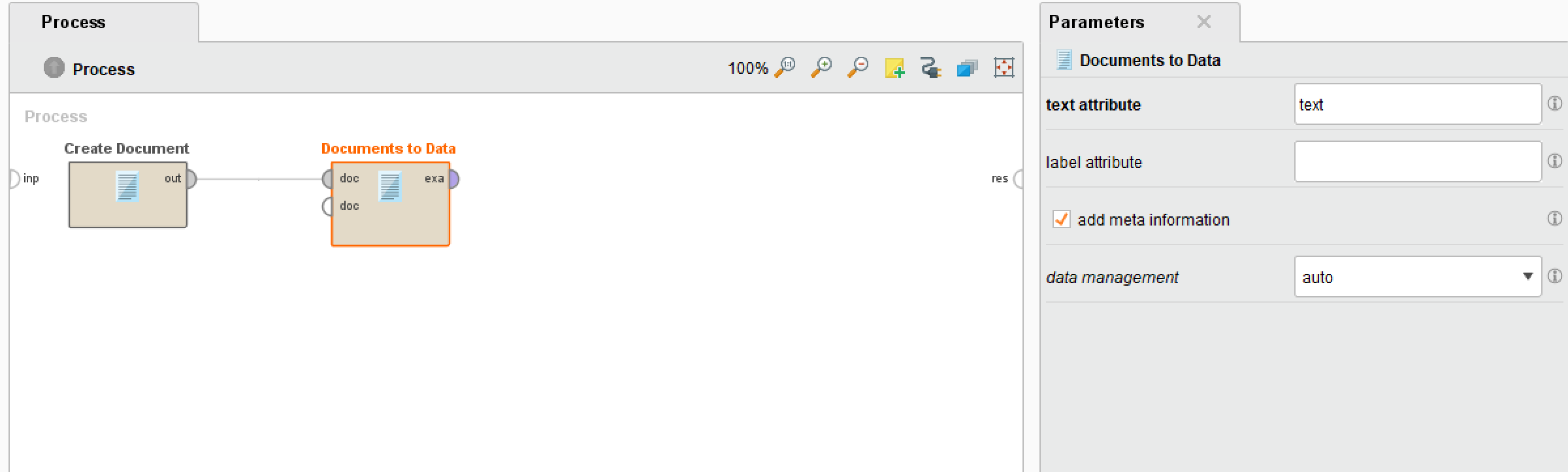
Create a new process, and add the Create Document Operator into the process window:



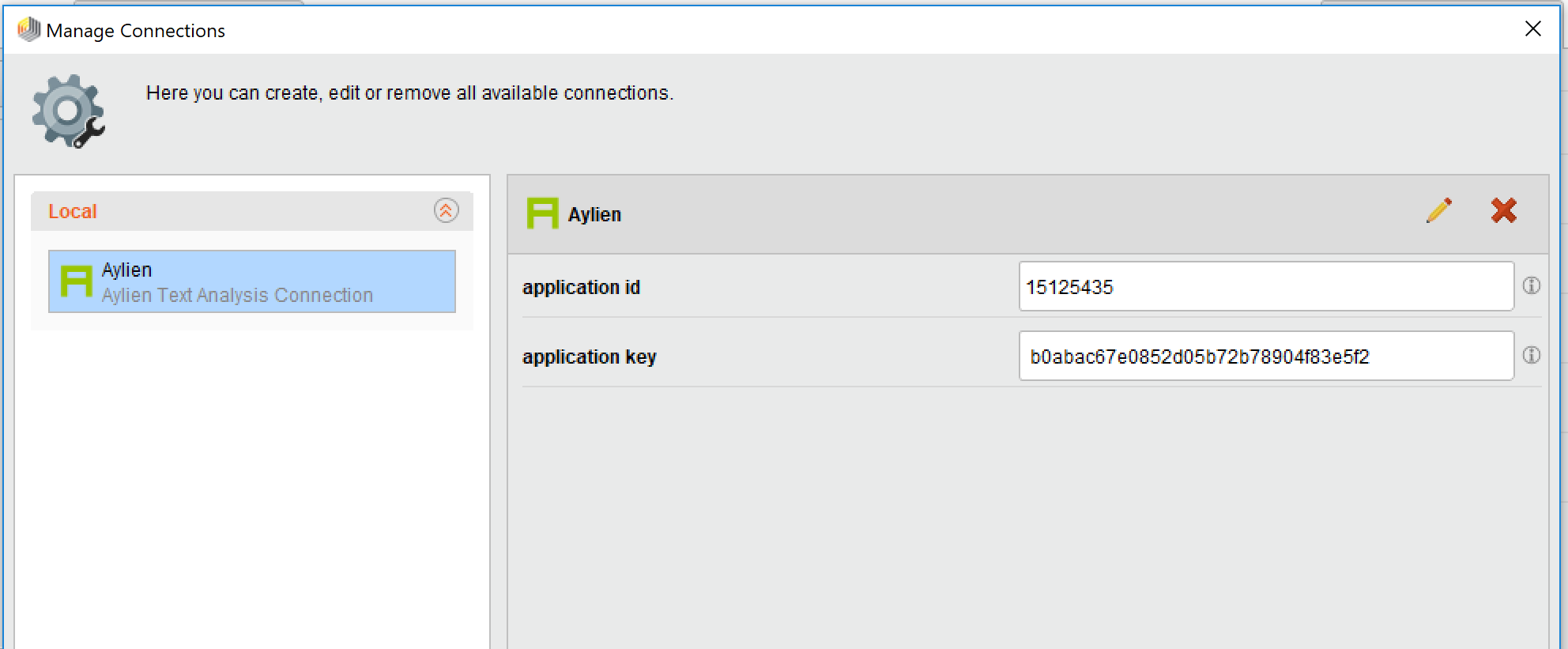
Next click in the Parameter window and enter a tweet from the tweets Excel file.



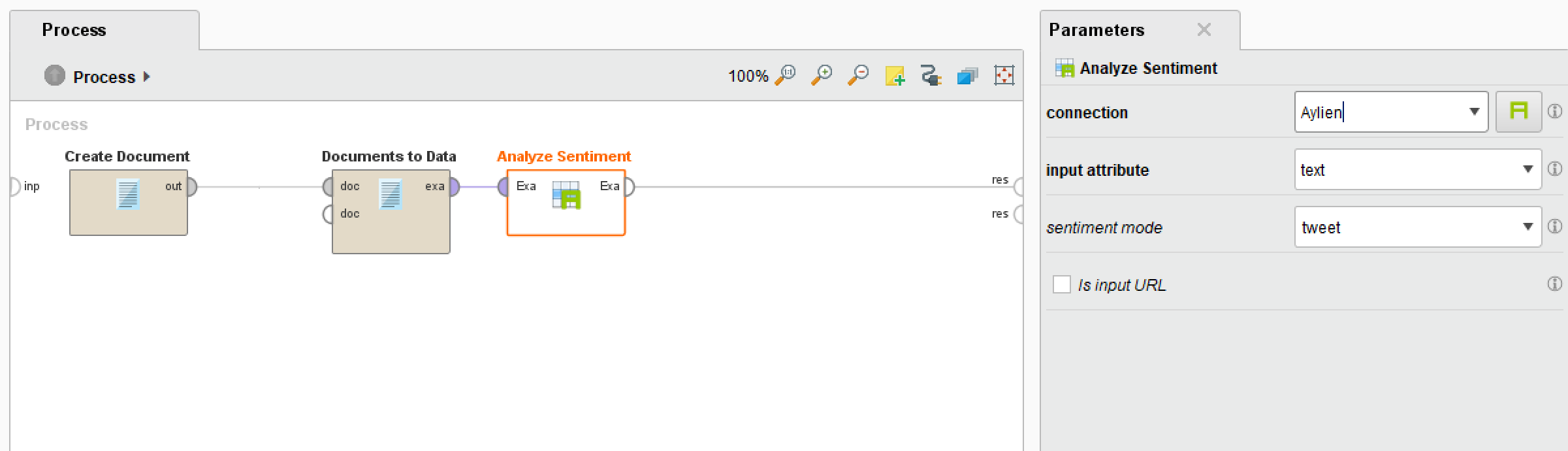
**Step 2:** Add the Documents to Data Processor



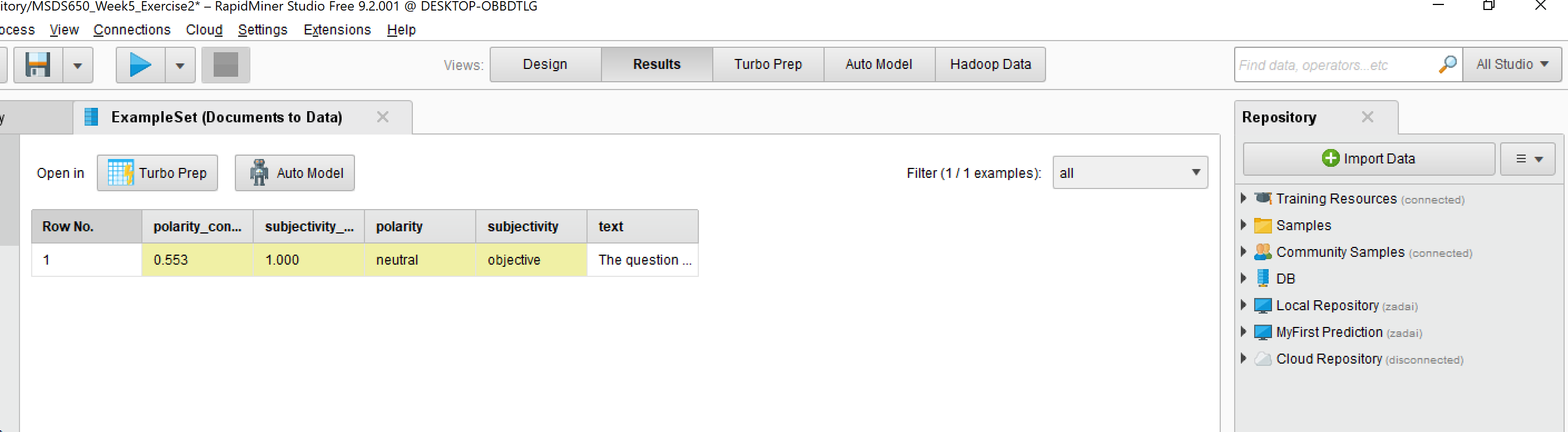
**Step 3: Setup your** **Text Analysis API (Application Program Interface keys.**



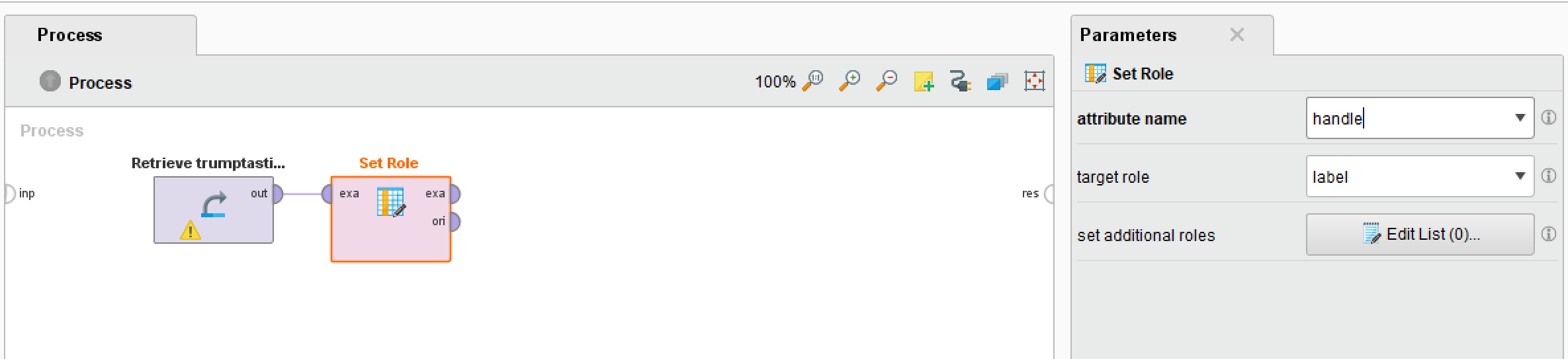
Add an Analyze Sentiment operator from the Aylien extension, and change **connection** to Aylien (or whatever you named your Aylien connection as), and **input attribute** as **text.**



Now click and run. I used the first tweet from Hillary and it came back as neutral.

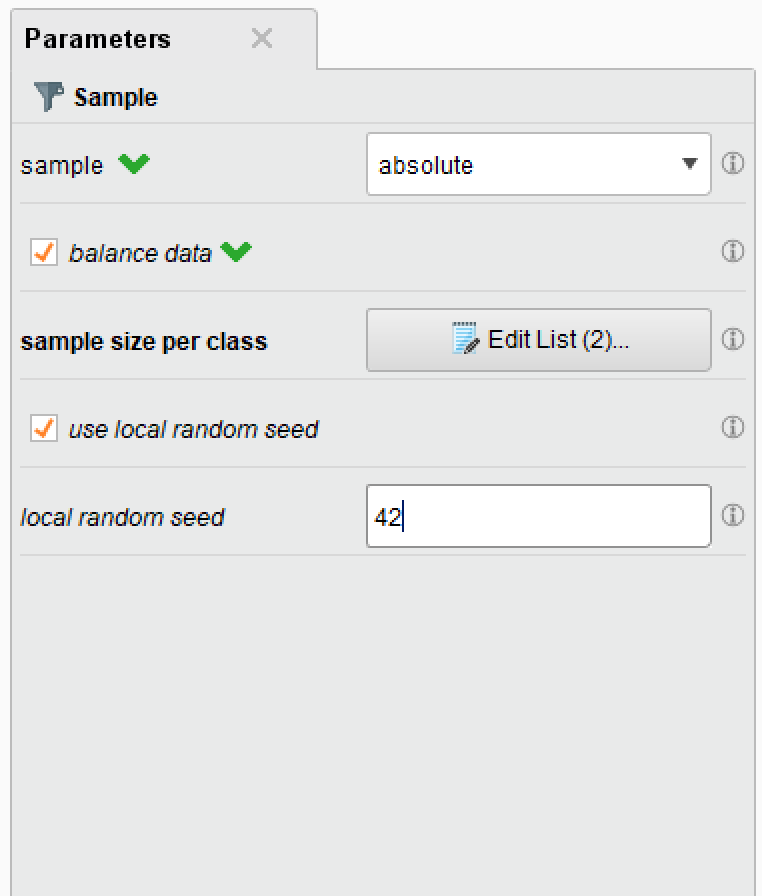


Now, I’ll go in and add the entire data file. I’ll upload the file using Add Data in the Repository. After I have the data retrieved I’ll connect it to a Set Role operator, then a Sample operator.

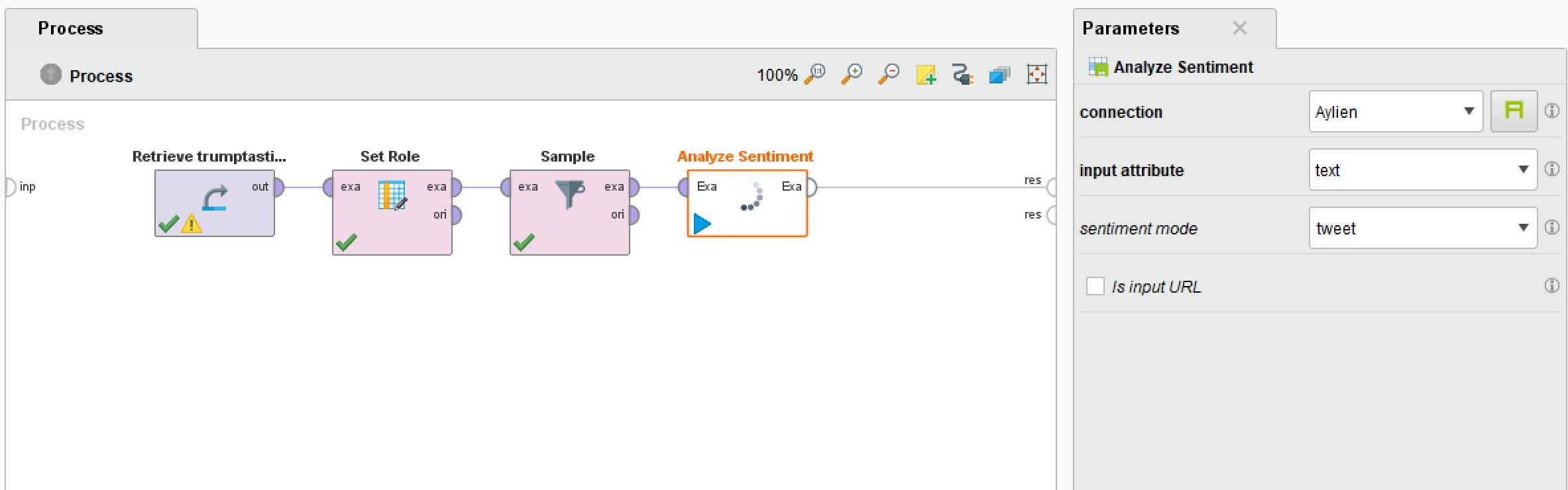


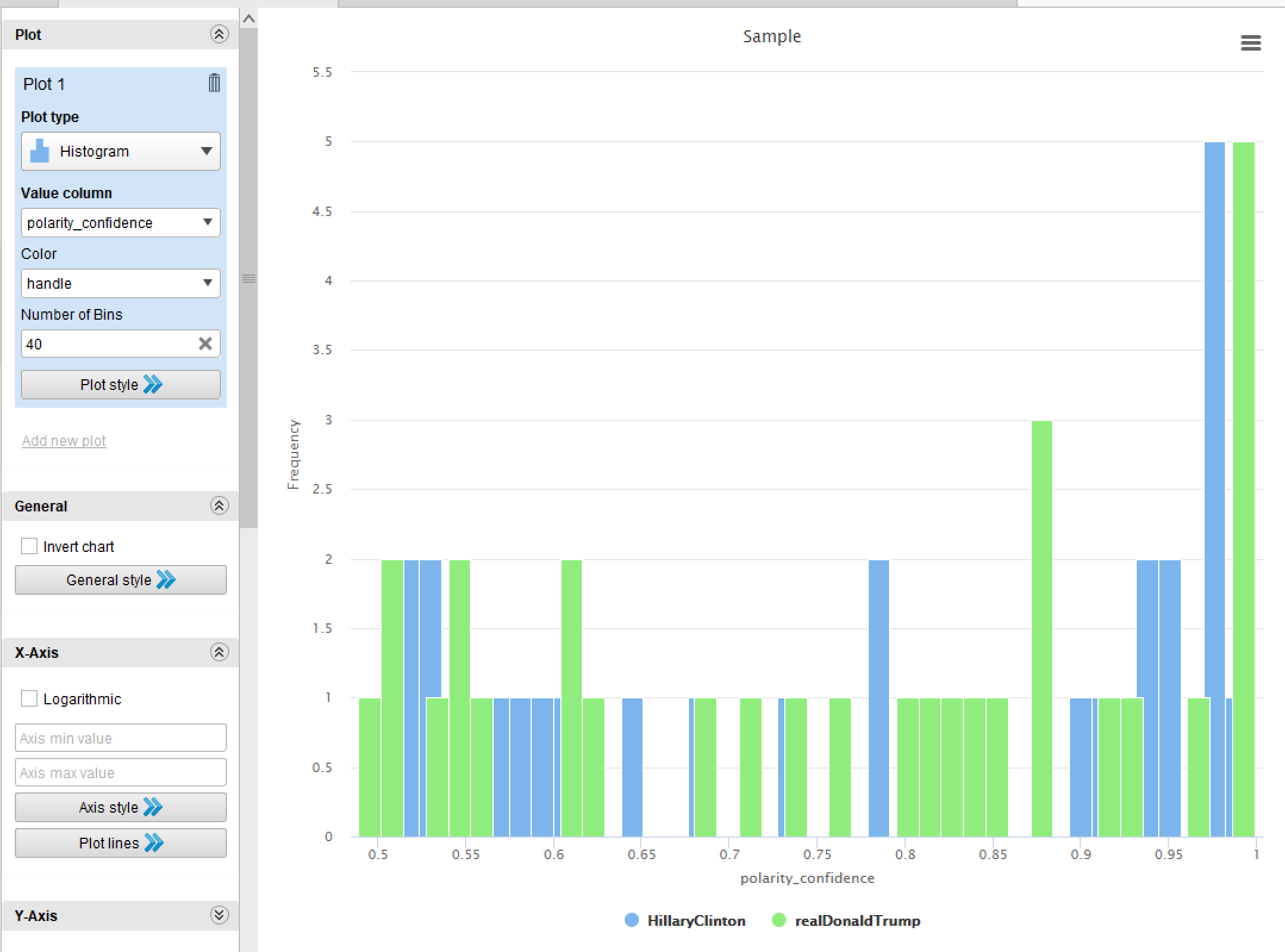


Now in the Sample operator’s parameters I will click on the ‘Show Advanced Parameters’, then check the box balance data, and click on ‘Edit List’ for **sample size per class:**



Finally, I’ll connect the Analyze Sentiment operator after Sample, using the Aylien **connection** and **text** as the **input attribute**.





It’s only 60 tweets so it’s not an entirely accurate representation of the more than 6000 total tweets but it gives us a good idea of the polarity of tweets between the two candidates of the 2016 election.

# Assignment Write Up for Week 5

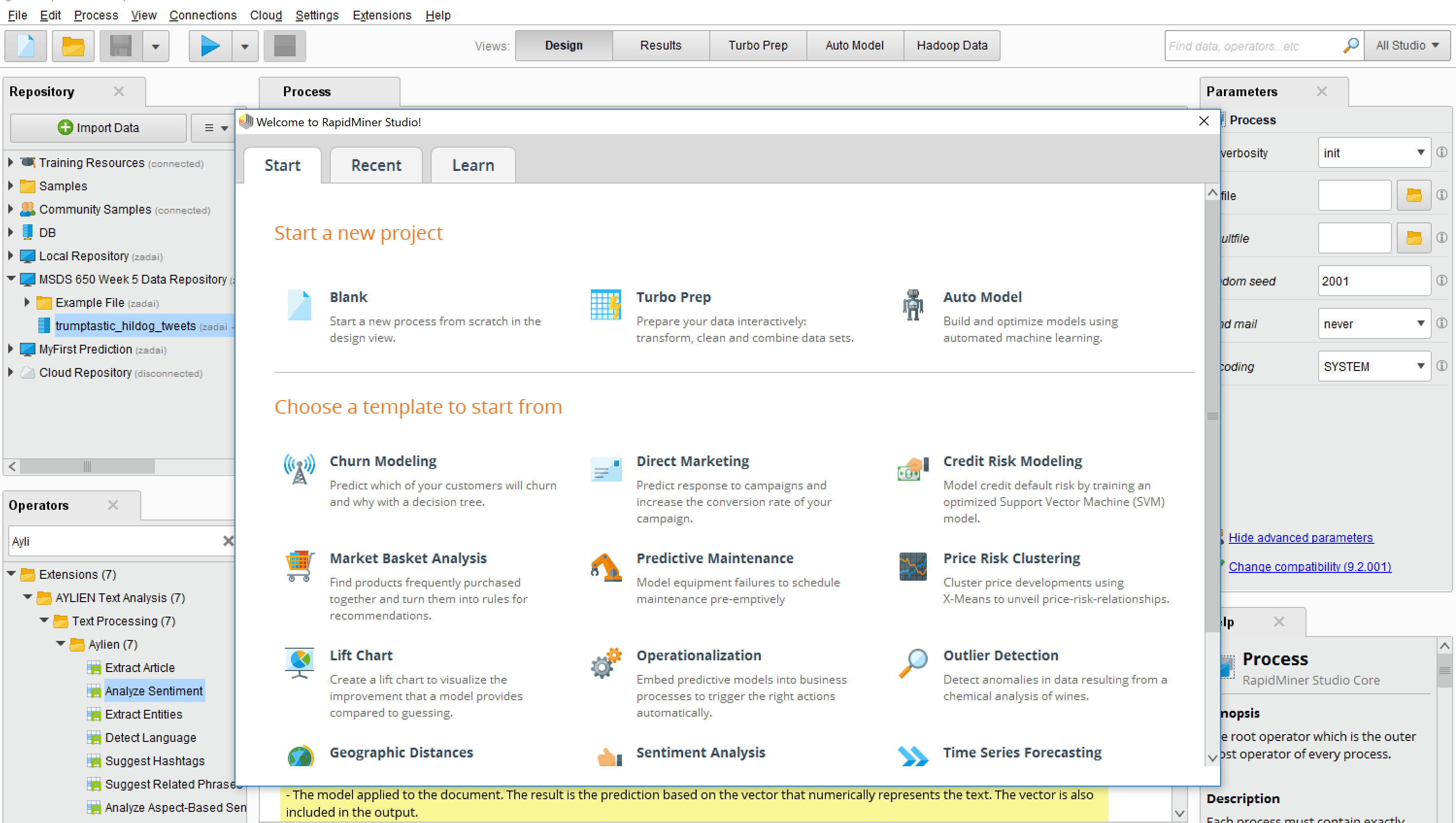
With the assignment for this week I will be using what I learned from the first two exercises as well as the Rapid Miner tutorial videos and doing a text processing and sentiment analysis with the application. I have decided to use one of the .csv files given to me, so before I go into rapid miner I’m going to open up the file and look at what is in the fields.

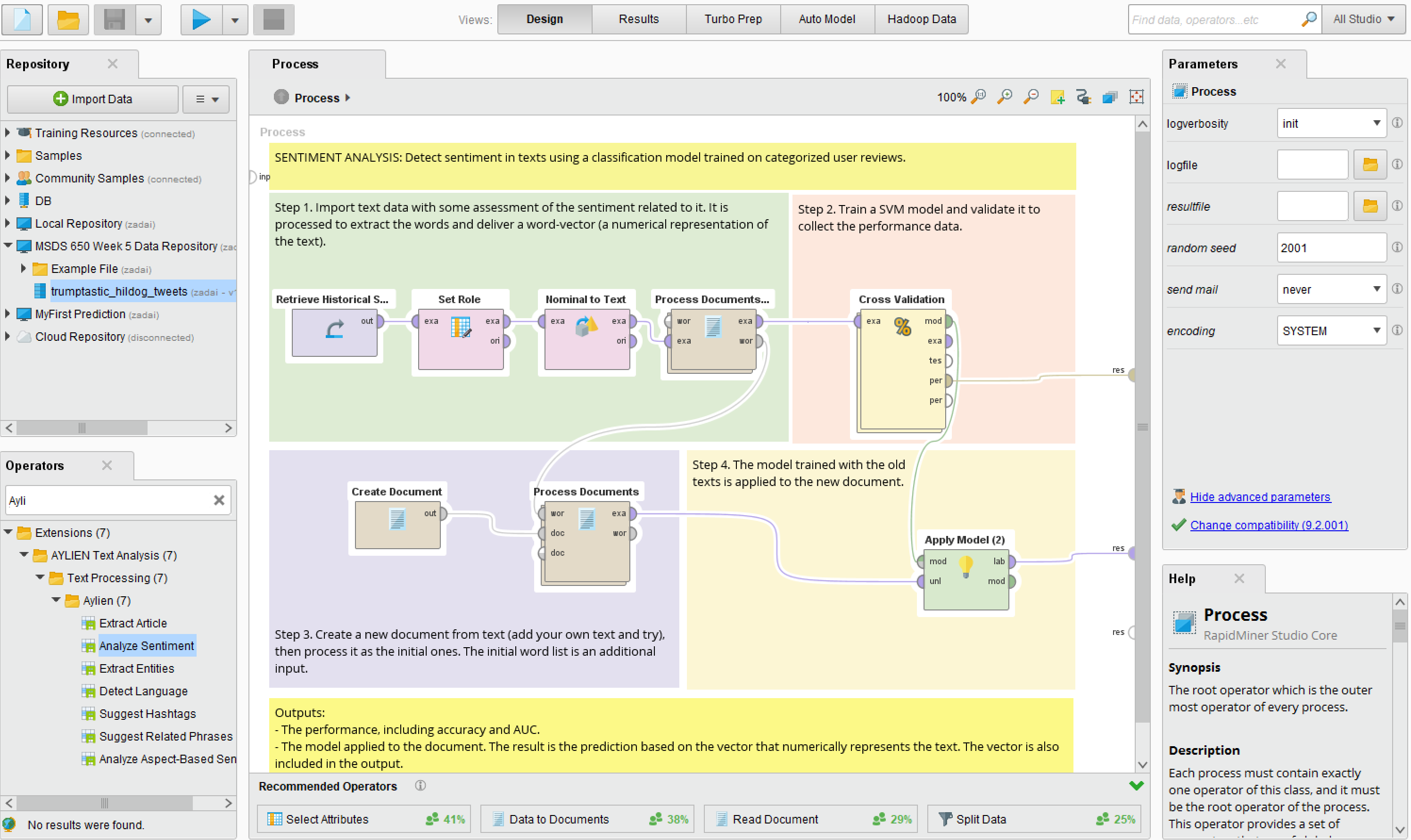
The file I’ve chose to analysis is for the Data Science skills:



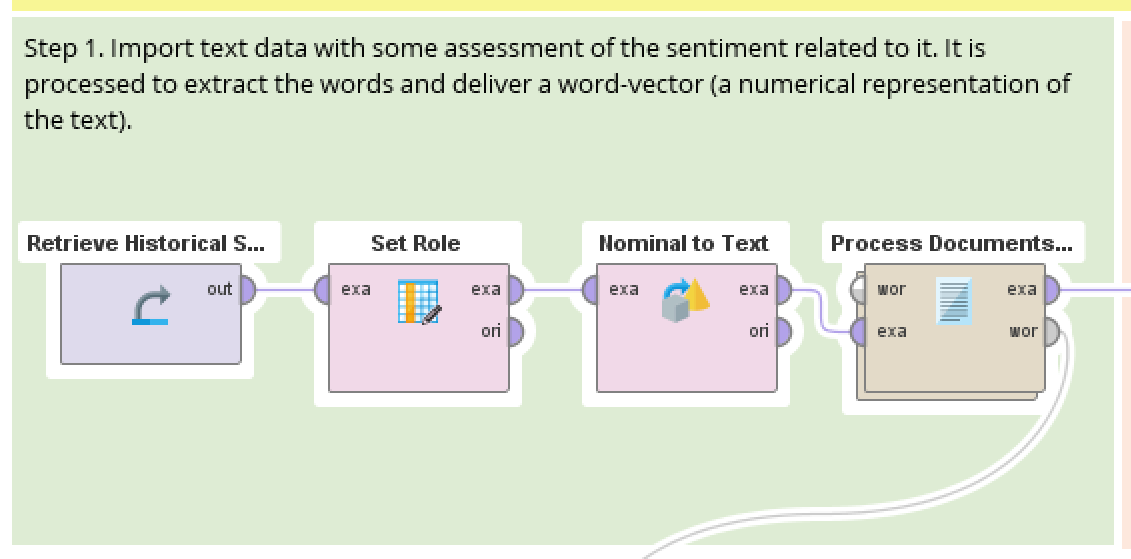
This data came from LinkedIn and it looks like job postings for open Data Scientist positions and the skills they require for their jobs. There is four categories of skill, cloud software, database software, statistical software, and programming software.

I will start by opening a new workspace in Rapid Miner, it gives you a template option with one being a Sentiment Analysis, so I am going to click on that and see where that takes me.

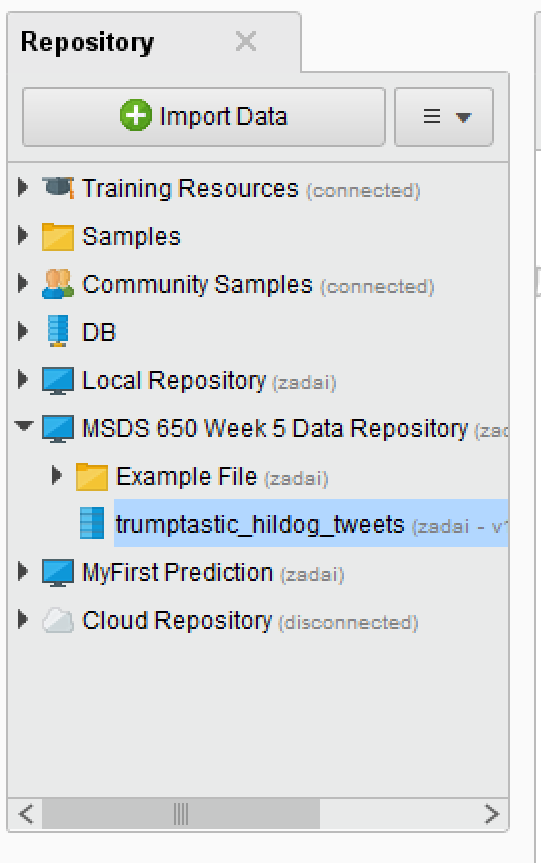




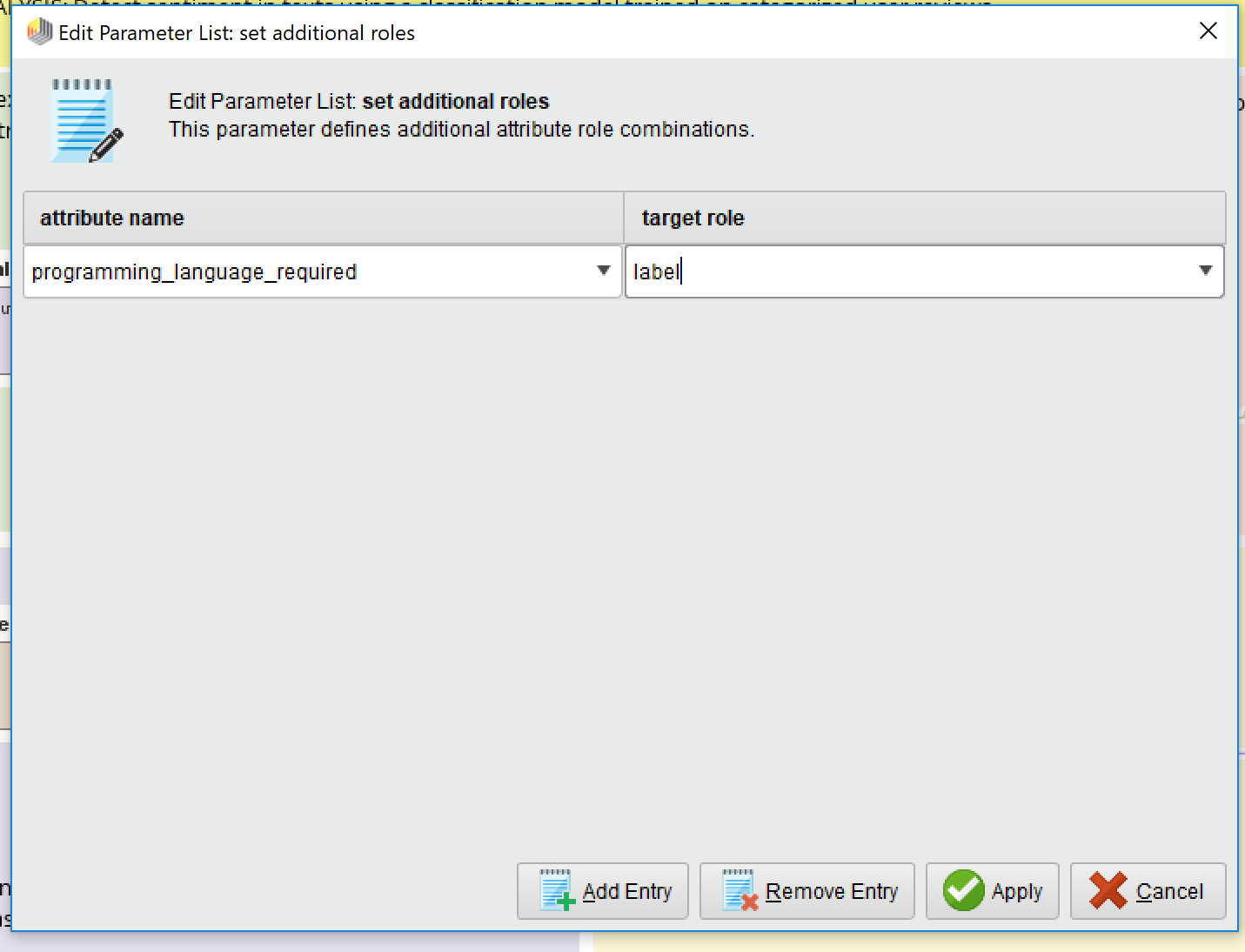
Starting from the top in step 1, where I will need to start by importing the data and then running it through several processes in order to analyze the data.



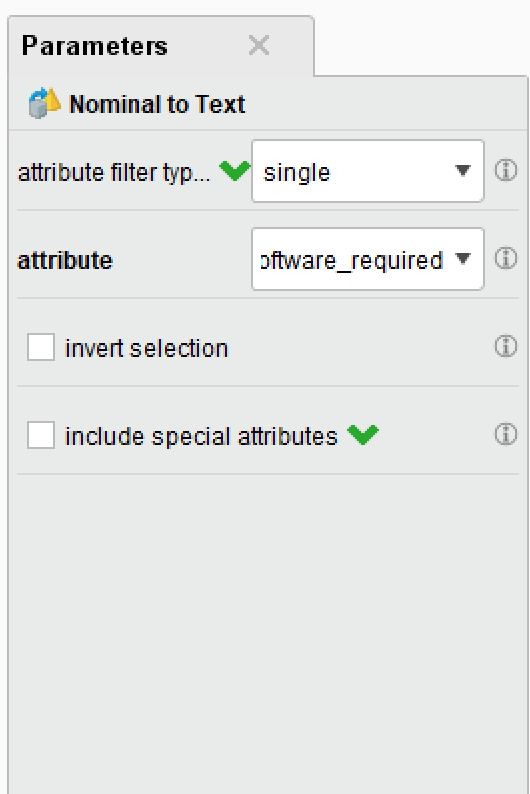
To import my desired file into the Retrieve data section I will first need to import data into my desired spot.



After importing the data and retrieving it the next step in this mapped out process is to set additional roles. For under attribute name I set it equal to **programming\_language\_required** and target role to **label.**

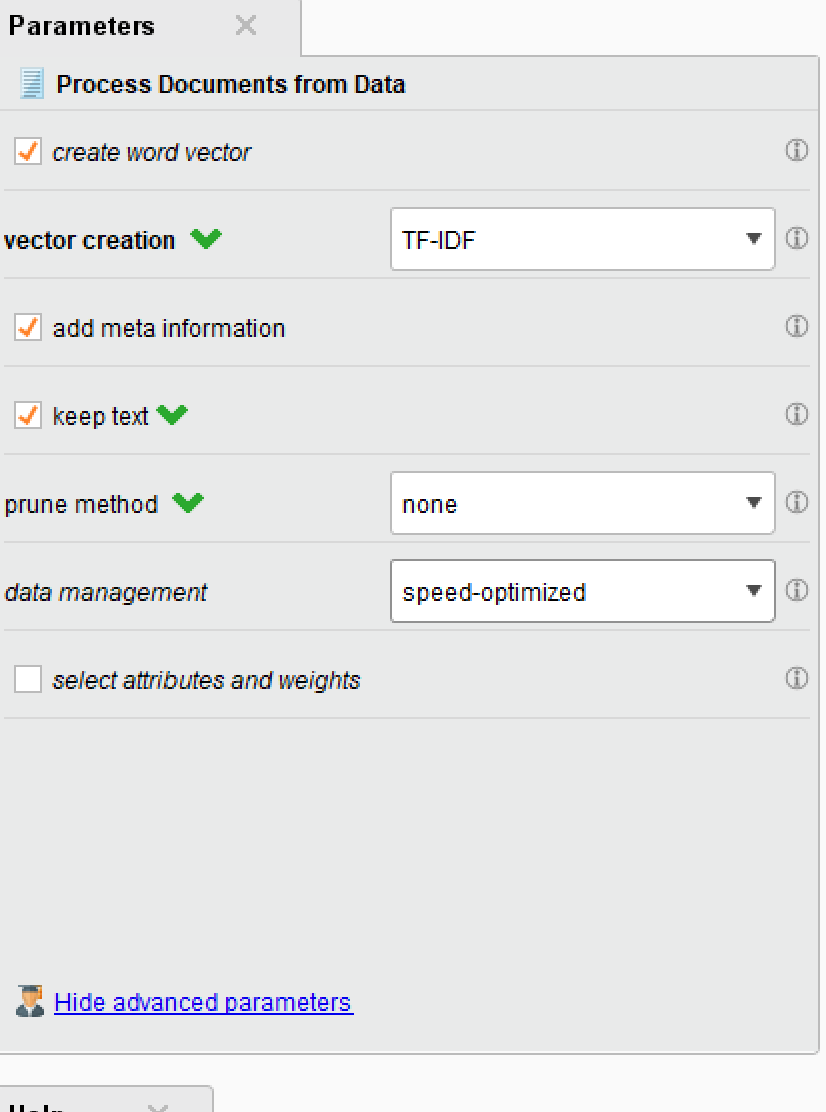


Next is the Nominal to Text process.

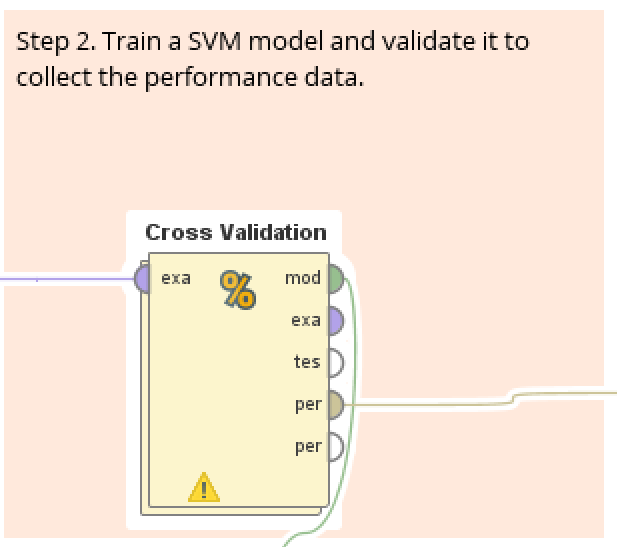


I set the attribute filter type to **single** and the attribute of choice is **statistc\_software\_required**

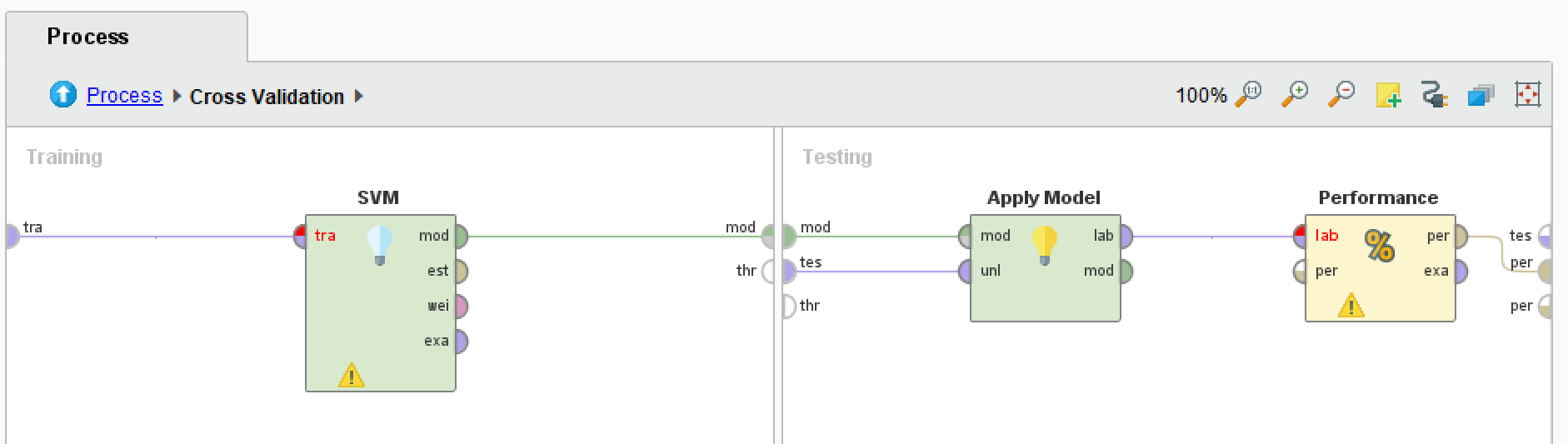
The last part of step one is Process Documents from Data.



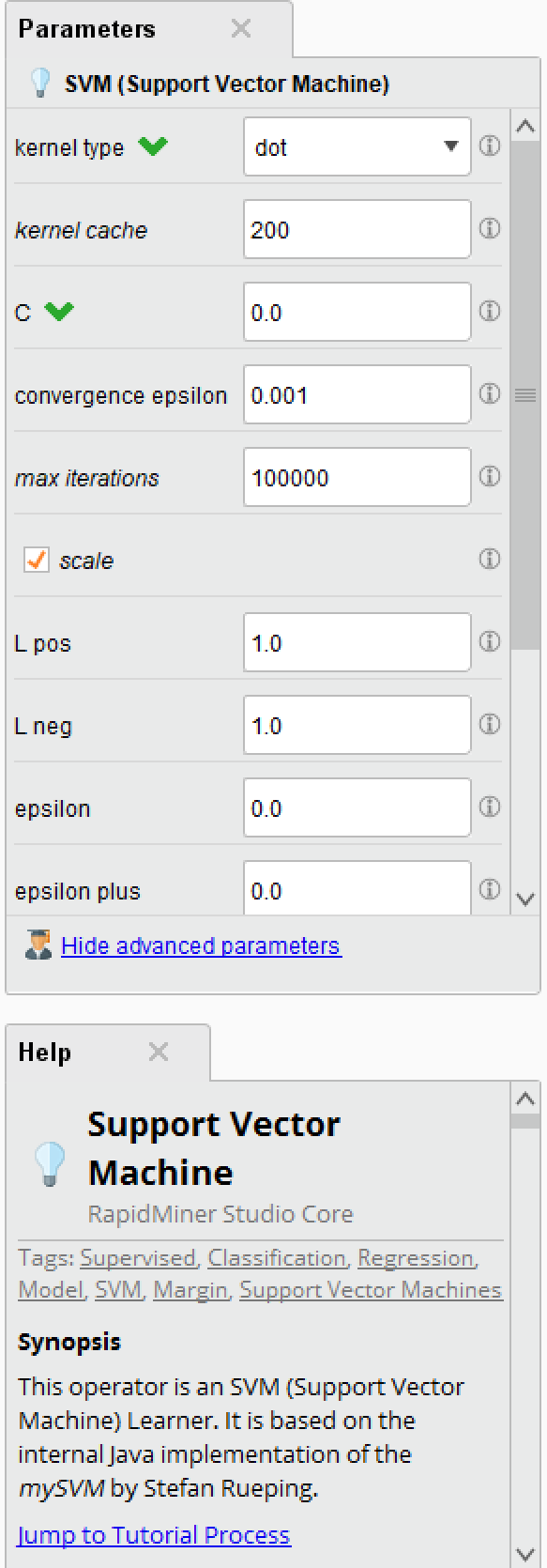
After this we are now onto Step 2, and in step 2 we are training the SVM model and validating it to collect the performance data.



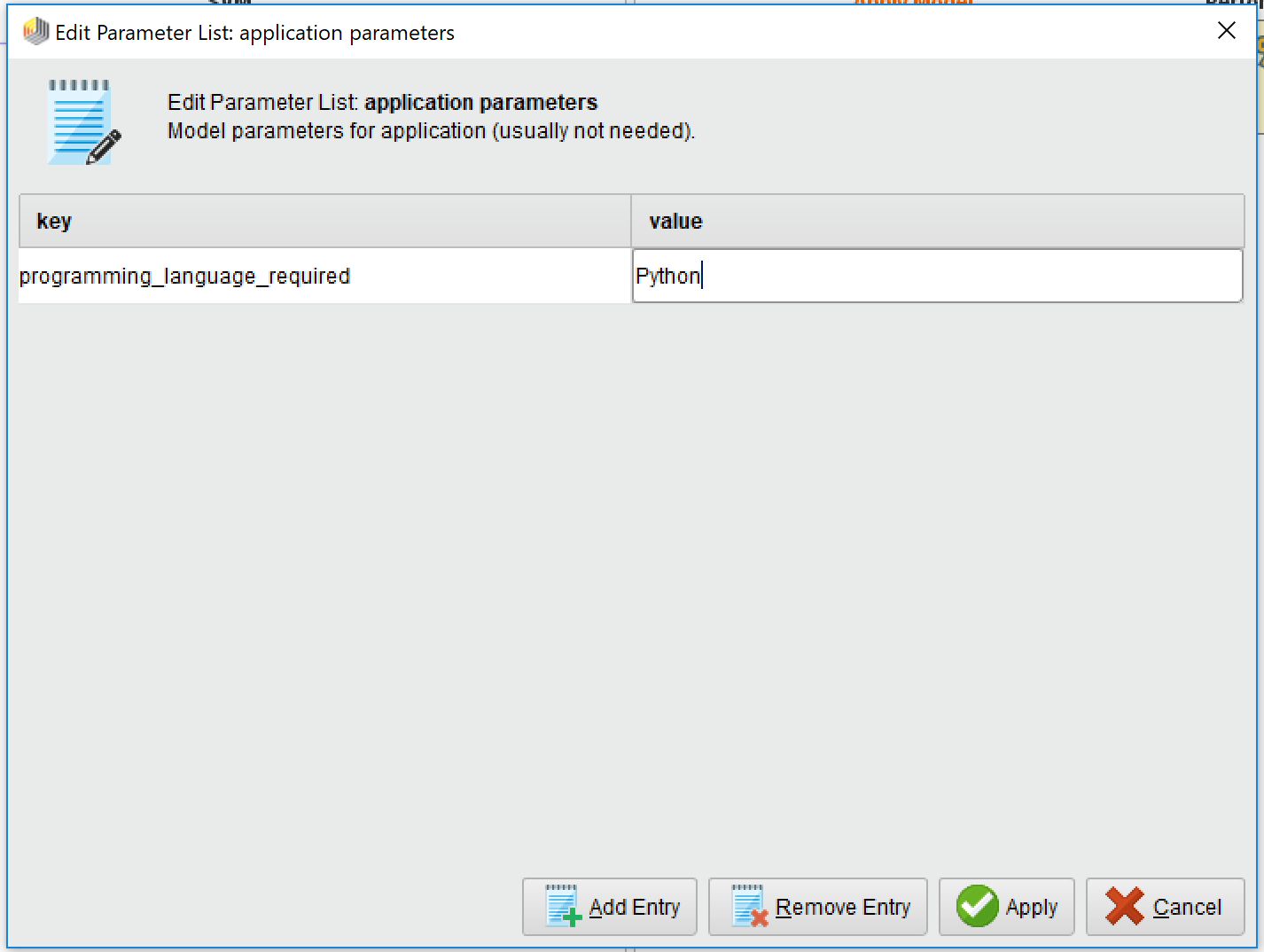
Clicking on the operator will allow us to dive in deeper to what the cross validation is doing.



Checking out each component of the Cross Validation I start in the training portion with the SVM.

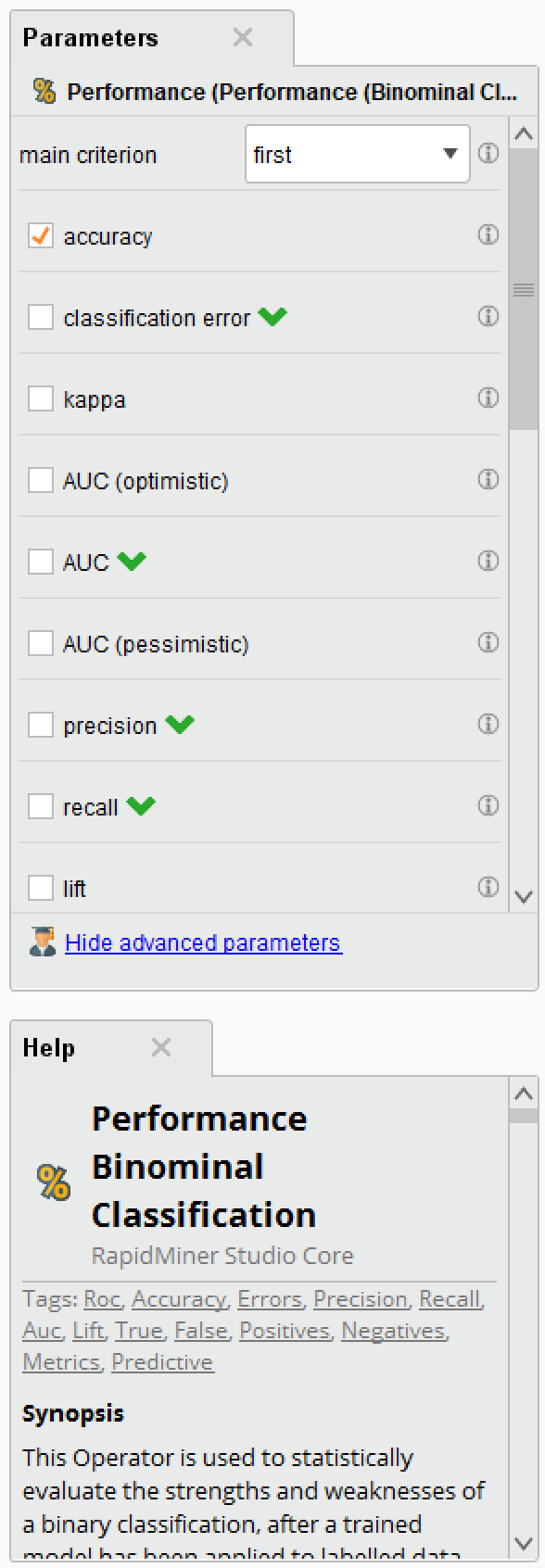


It is cool to see the intricate parts of the SVM as well as get a brief little overview of what an SVM is.



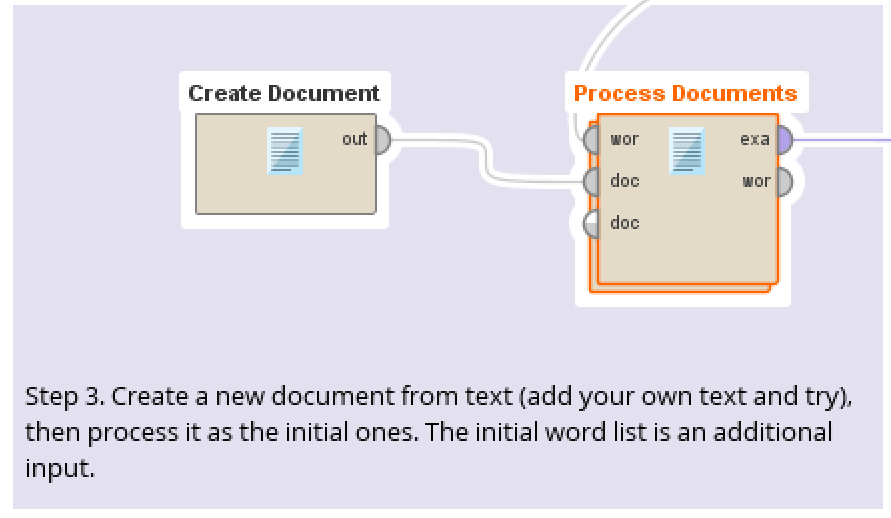
For the apply model, I decided to add something to the parameter list, I added my key as **programming\_language\_required** and he value I set to **Python**

And last I looked at the performance operator:



Here were able to look at the different parameters that encompass the performance of the cross validation process.

Next, I’ll look into step 3, where we will create a document and with the help of our result from step 1, process the documents.

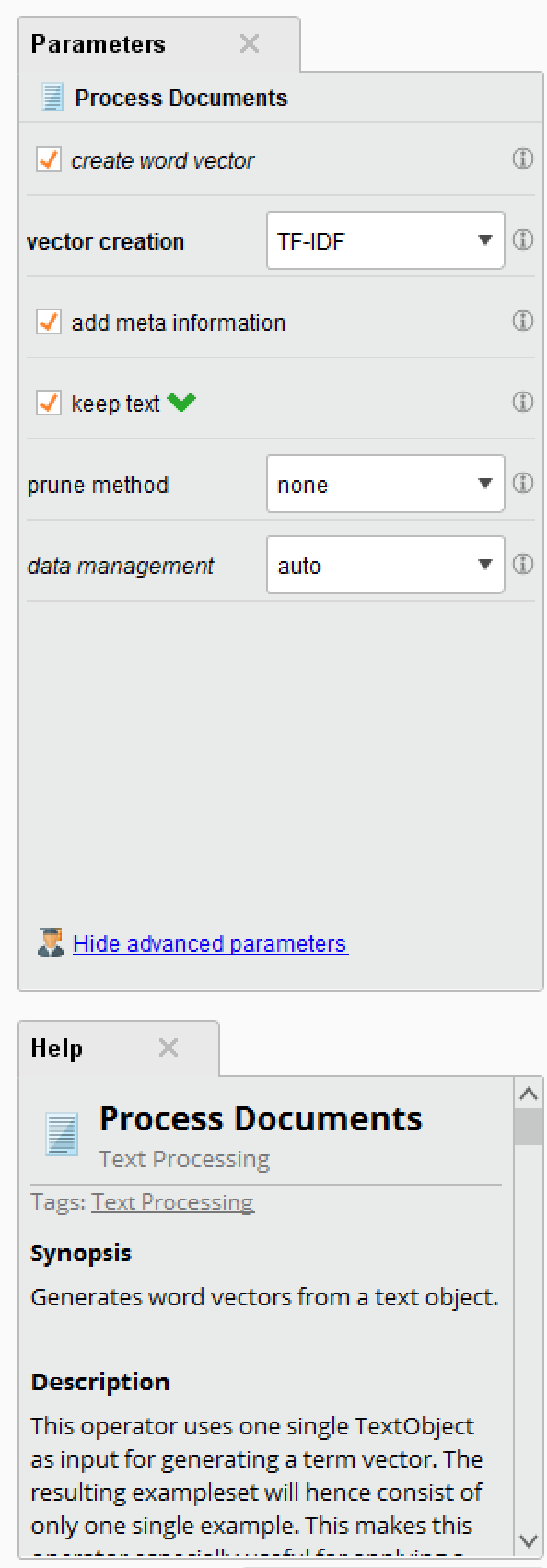


I looked at the Create Document operator and looked into the Edit Parameter Text, this is the text I found.



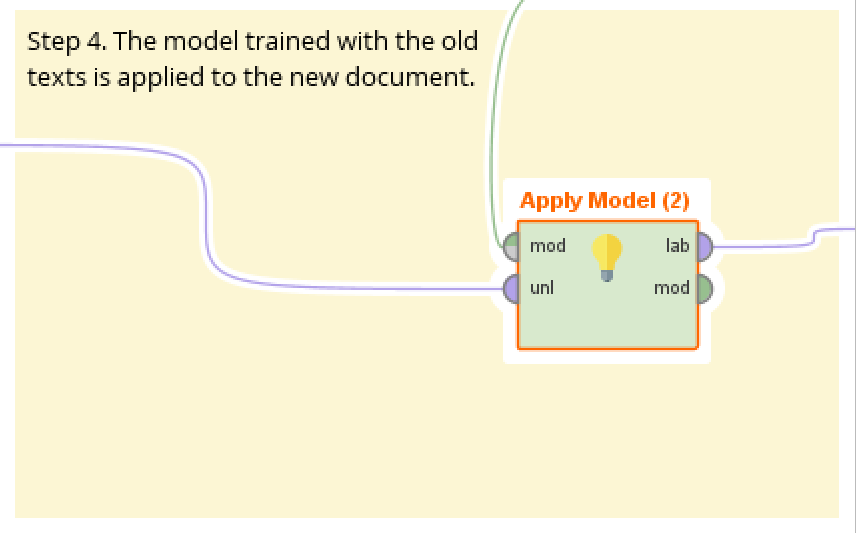
It appears to be some sort of short story, I’m not entirely certain but I found it an entertaining read nonetheless.

Now I’m taking a look at the Process Documents operator:



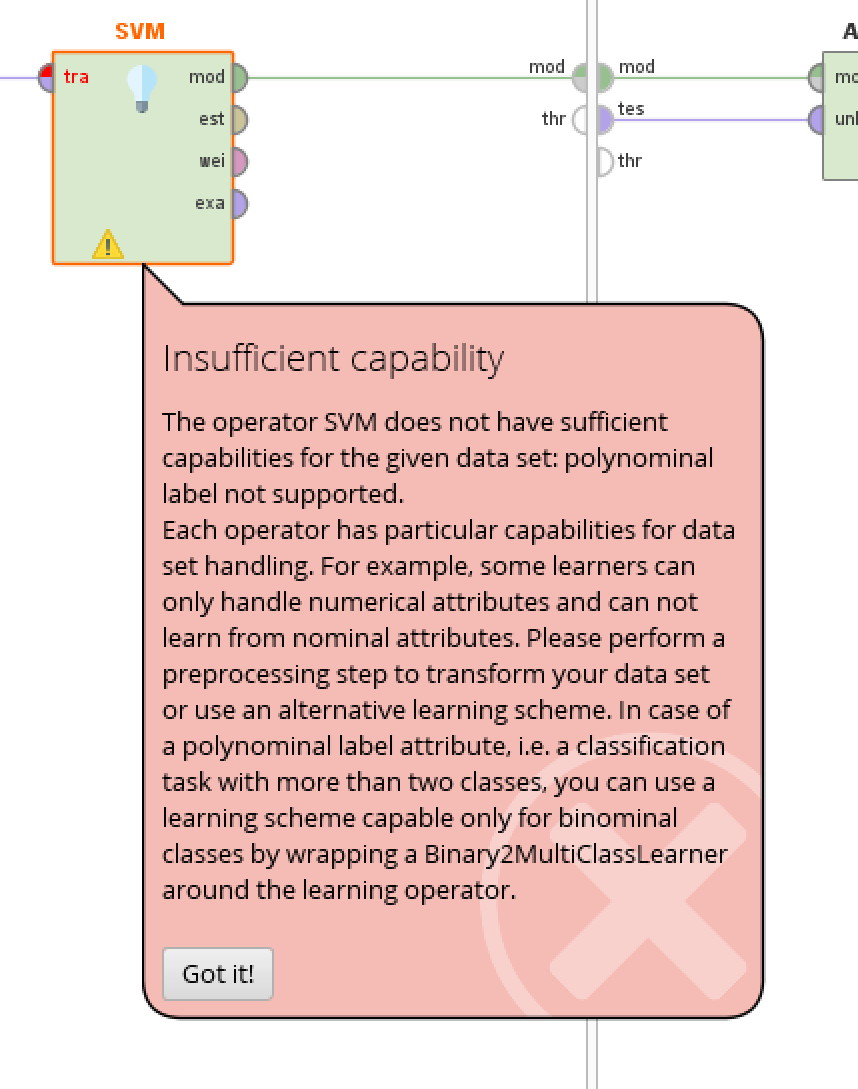
This operator also had a help section which describes how the documents are processed.

Finally I’m at step 4 which will apply the whole blending the old and new text. Since the two texts are very different I’m interested to see how this will turn out.



I’ve ran into a multitude of errors going through the process, so since I’m having these issues I’ve decided to change the data that I retrieve in the very first step to match what my Create Document is like in step 3.

The data I am now retrieving for the Sentiment Analysis is my document of Aesop’s Fables.



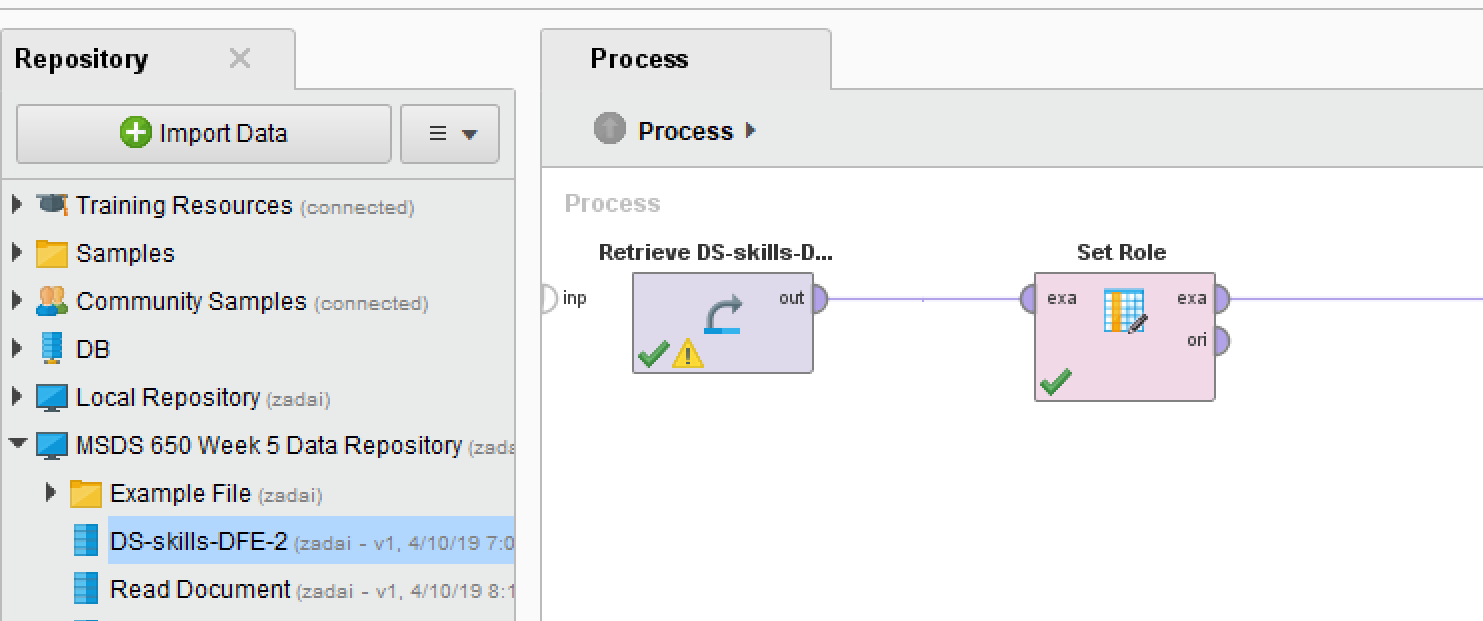
I’m running into the same issue that was bothering me before named Insufficient capability, I am unable to convert my text from a polynomial to a suitable data type for the model so I will have to scrap this and just freelance the assignment from here.

I will start again by using the original data file I proclaimed I would be using which is the Data Science job listings from LinkedIn.

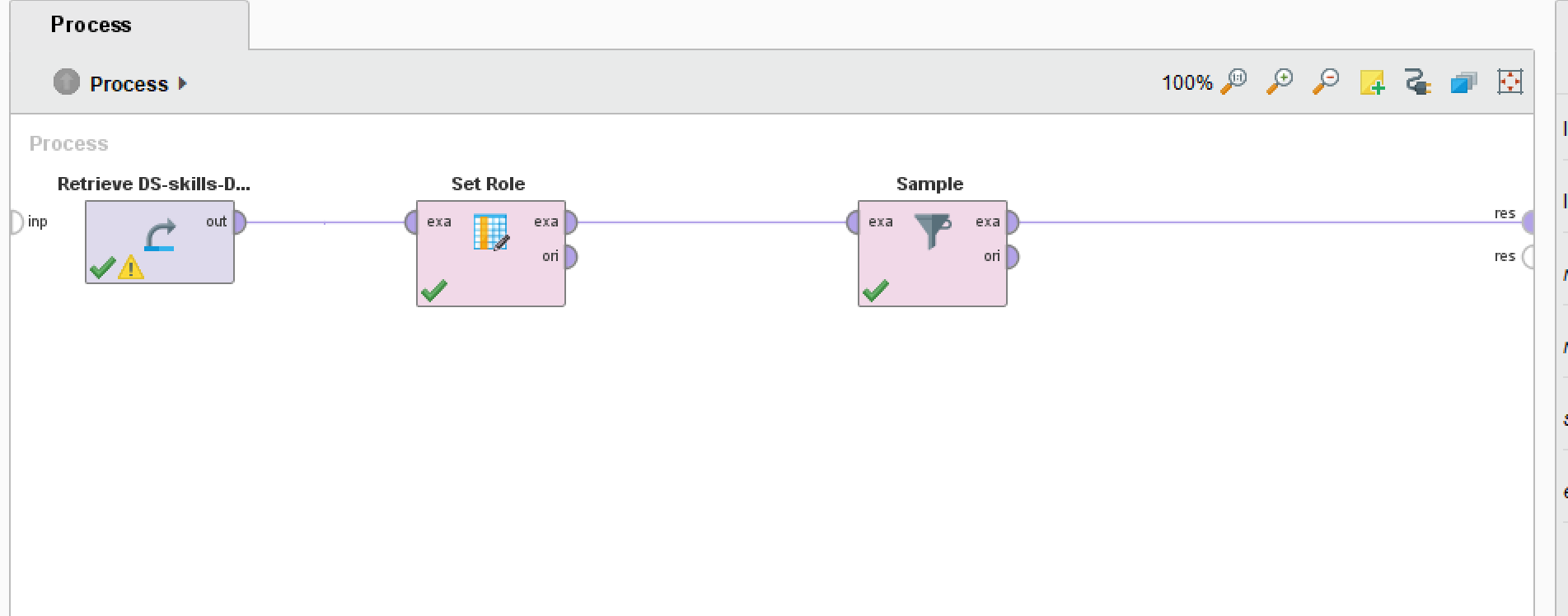
Since I already had it in my repository it just took a simple drag and drop into the process.

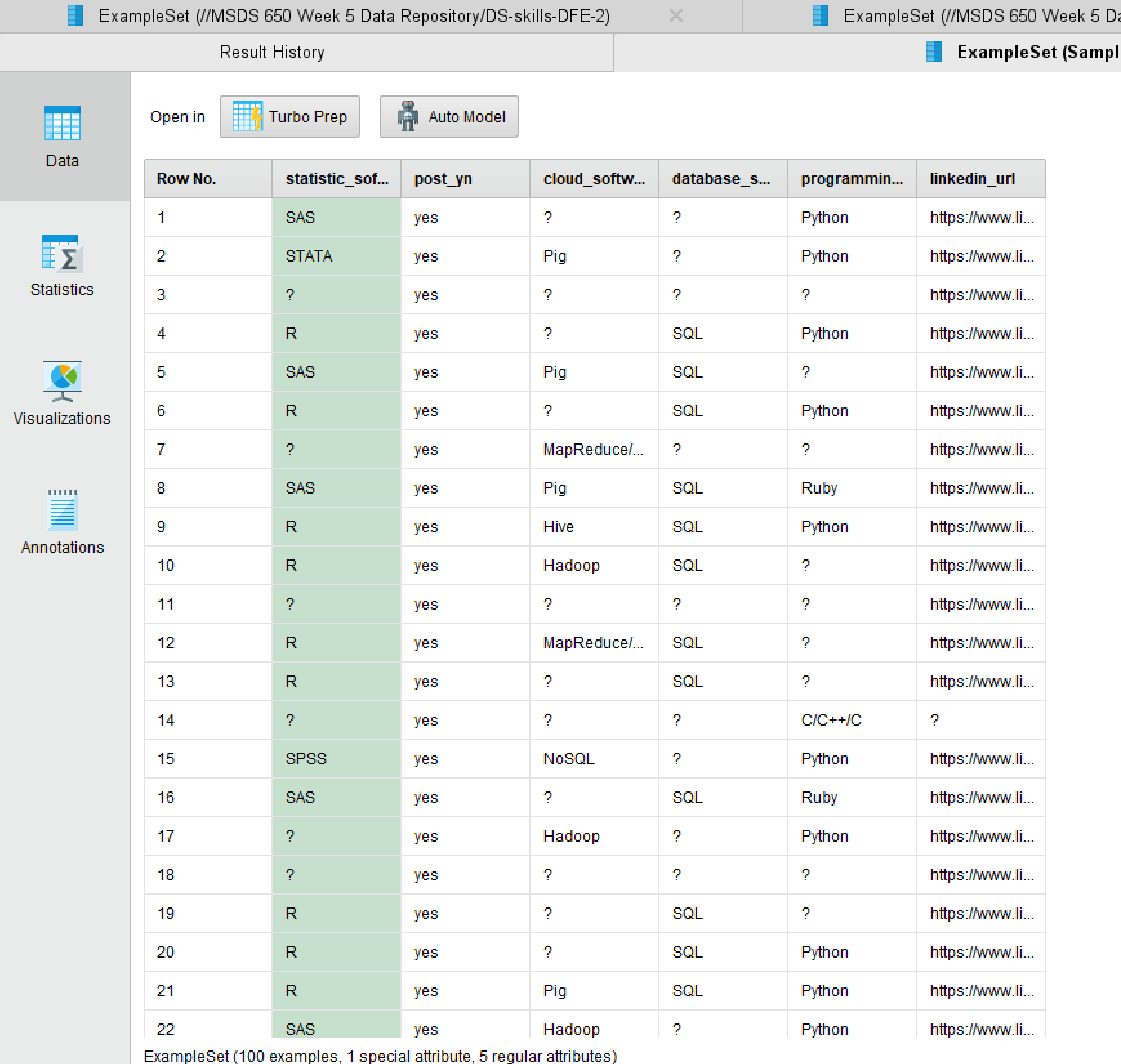


The type of analysis I would like to run on this data is a sentiment analysis. So we will run a Set Role to play around the different fields I can group the data by.

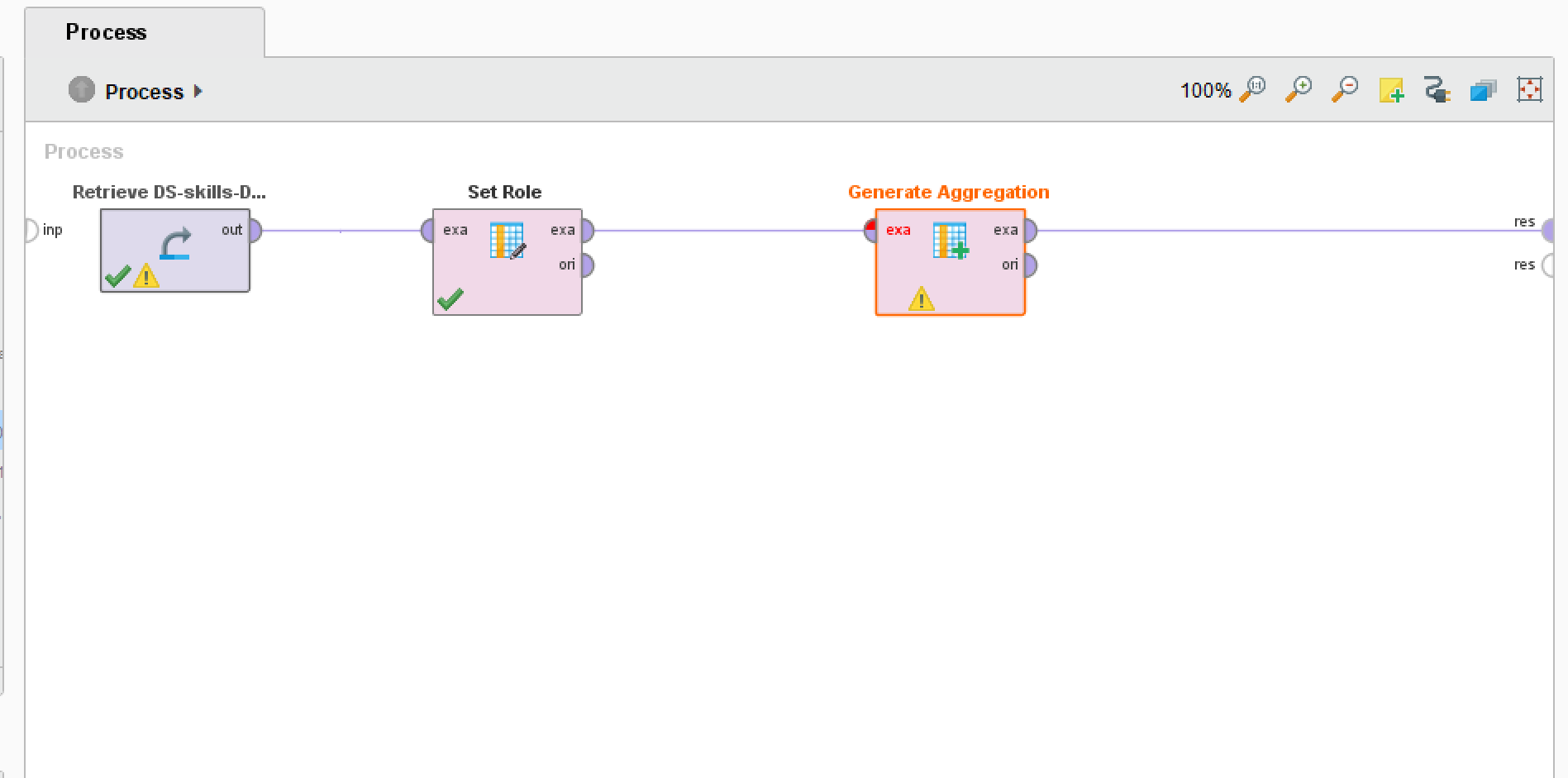


Then add in a sample operator to shrink the data into a more digestible bite.

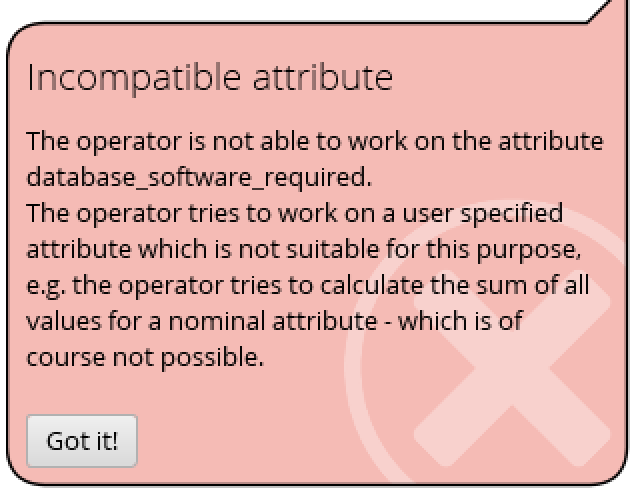




Next I brought in aggregate operators in hopes of aggregating one of the fields.

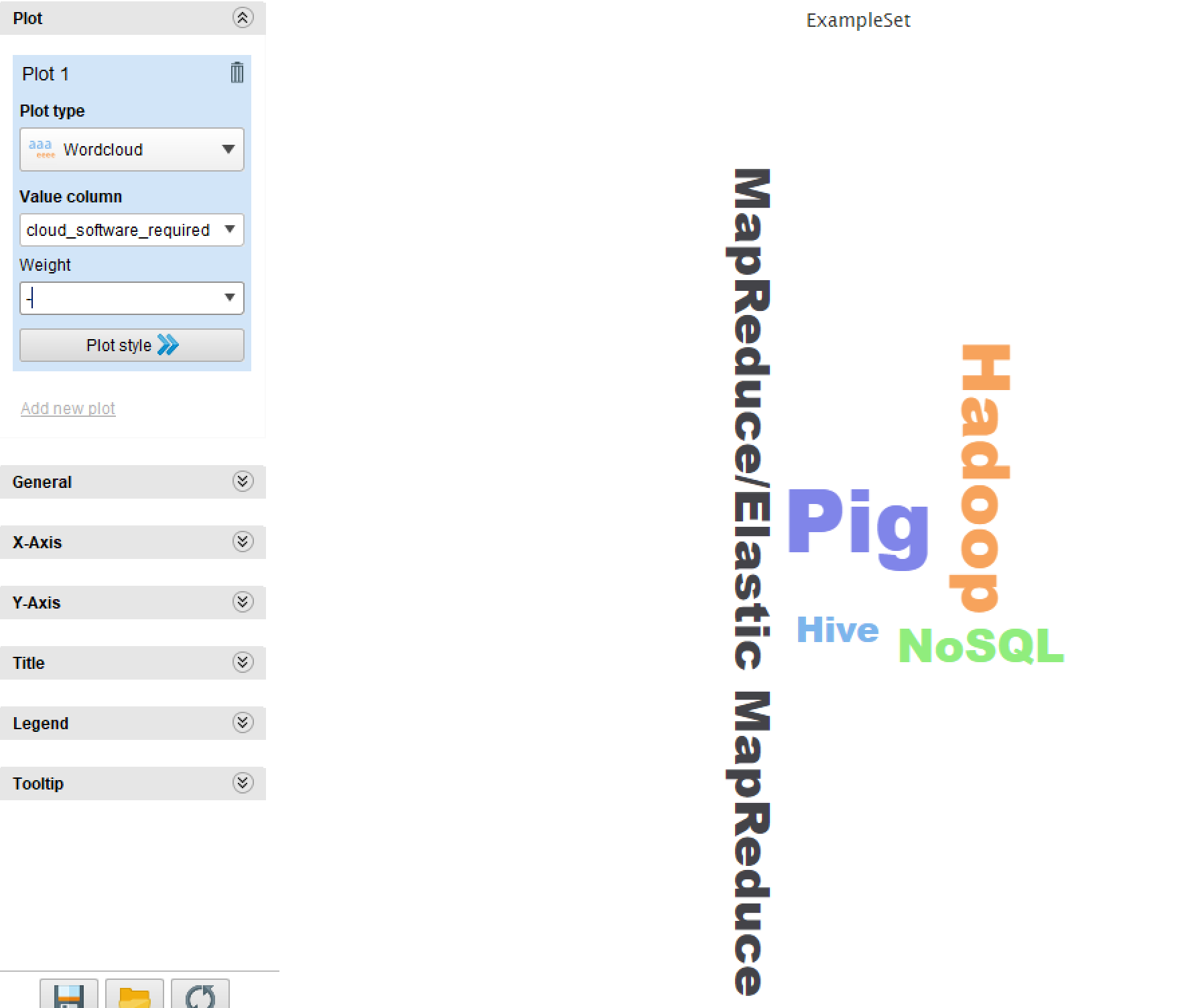


Like here for example, I tried aggregating the database software field, but was unsuccessful because it is nominal data.



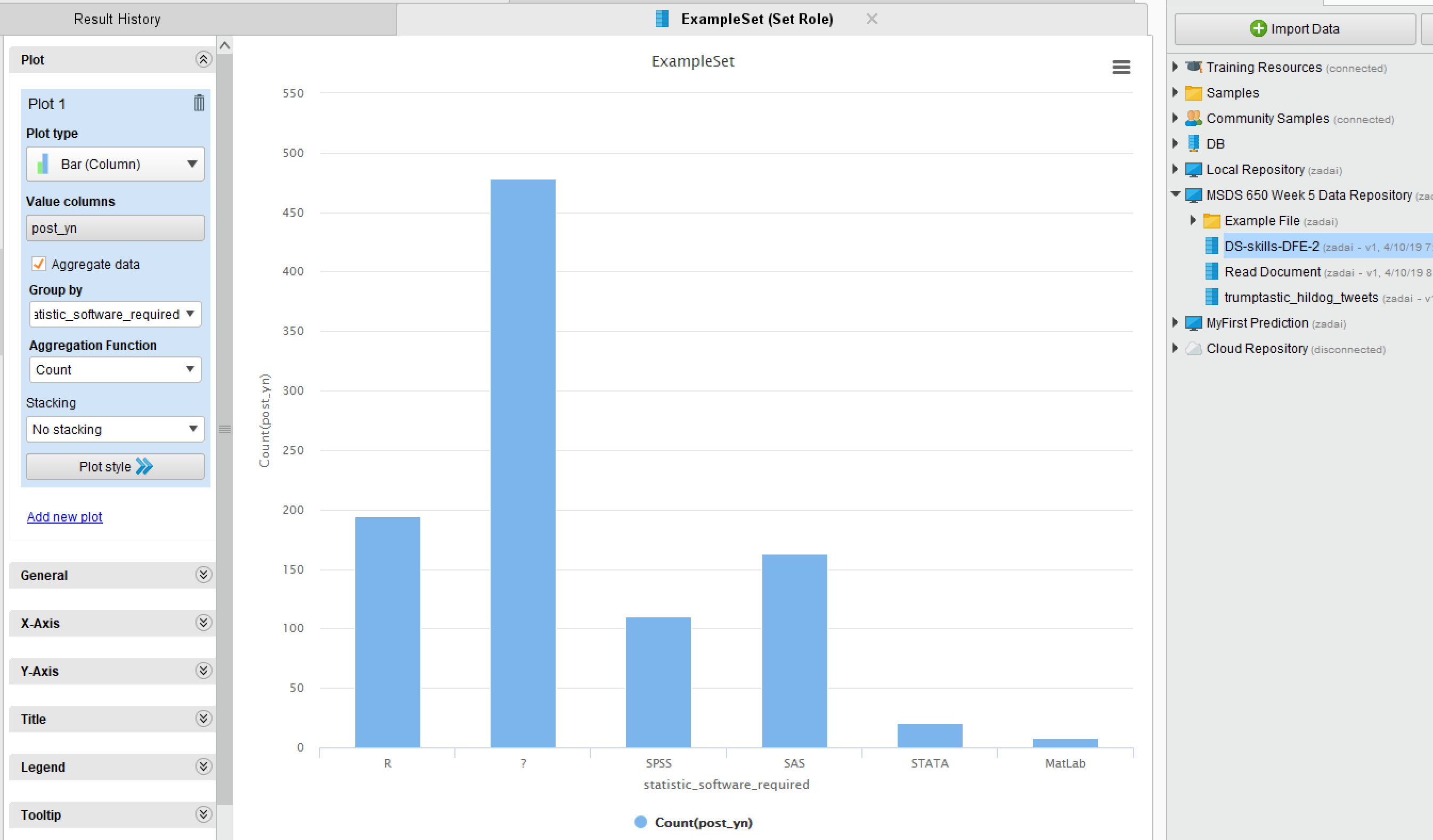
Moving over to the visualizations part of Rapid Miner I was able to play around with plots I’m familiar with but haven’t had a good chance to play around with in school or at my job.

The first one was a Wordcloud, which jumbles all the different possibilities of a field together and the size of the words is dictated on how many records they have.



From the looks of it you will notice not one stands out fully but Pig definitively looks like the biggest of the cloud software in the lay out and Hive looks to be the smallest. This weighs out to be accurate as well because when you run the cursor across the different words it will tell you what their sums are, here Pig was tallied at 123 and the highest while Hive was the lowest at 52.

Next I looked at bar charts comparing the volumes of statistic software and seeing how those stack up and some of the different things I can do to the graph.



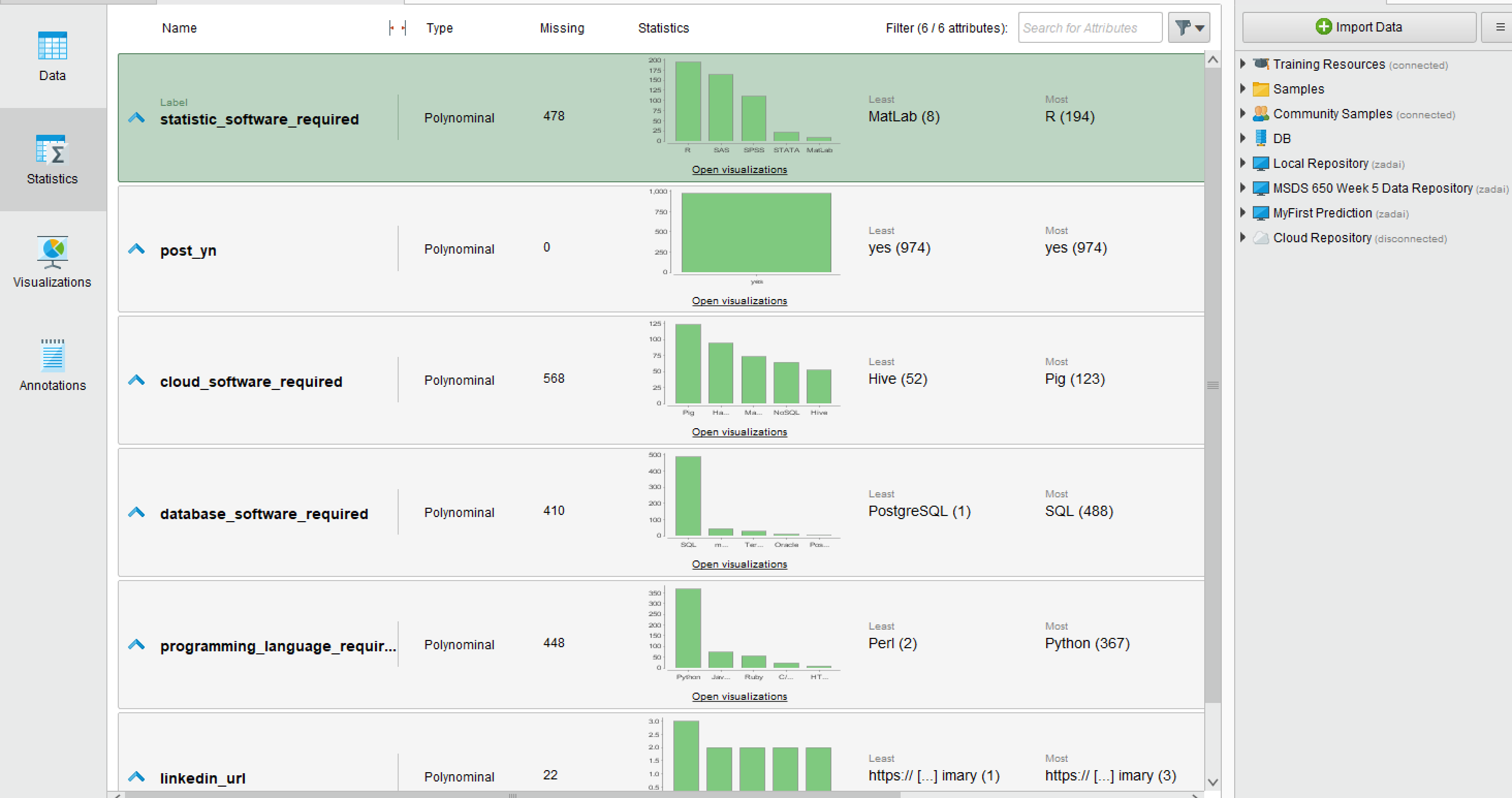
From our data it is overwhelming that ? or null is the highest when it comes to statistics software, out of the software’s R is the highest by volume and Matlab is the lowest.

Next I’m going to create a pie chart to map the different database software programs and see which one is asked for more on LinkedIn.



From above it is obvious that what is offered most likely is looking for people who know SQL or null again is very high at second. I read that as, with this job posting you don’t need to know SQL that well if it is the 2nd biggest value in the chart. But when it is null then we would think either their record is incorrect or the job doesn’t require the skill of knowing database softwares.

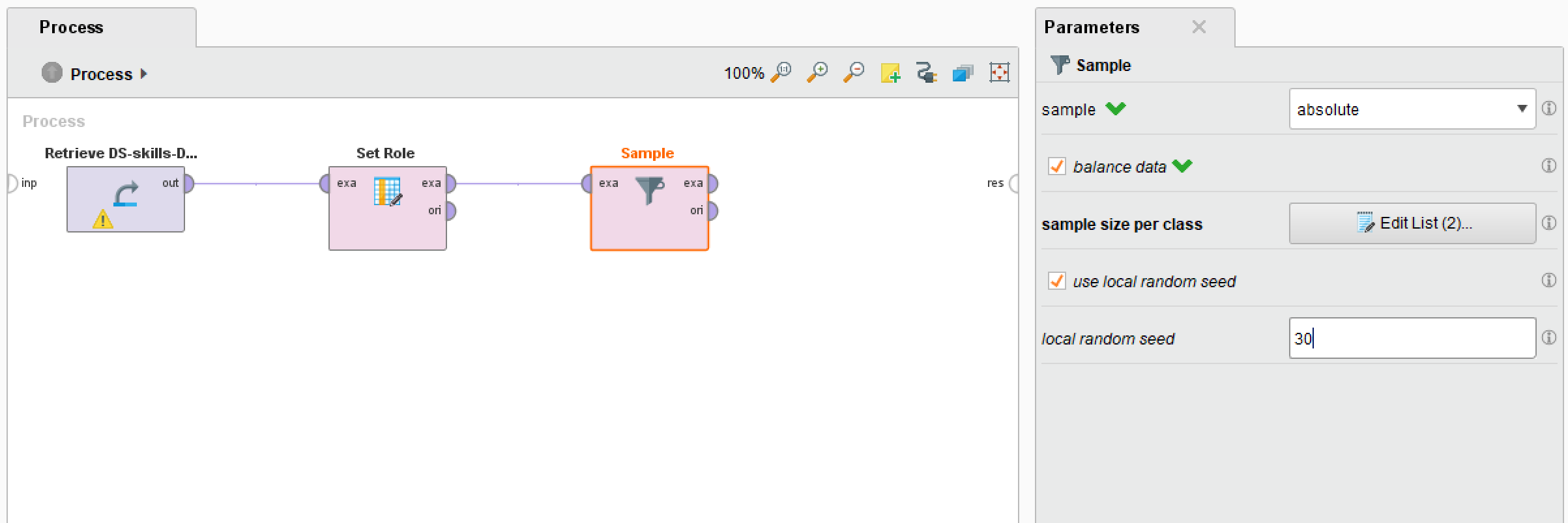
A cool thing about Rapid Miner is you can also see the summary statistics of the data



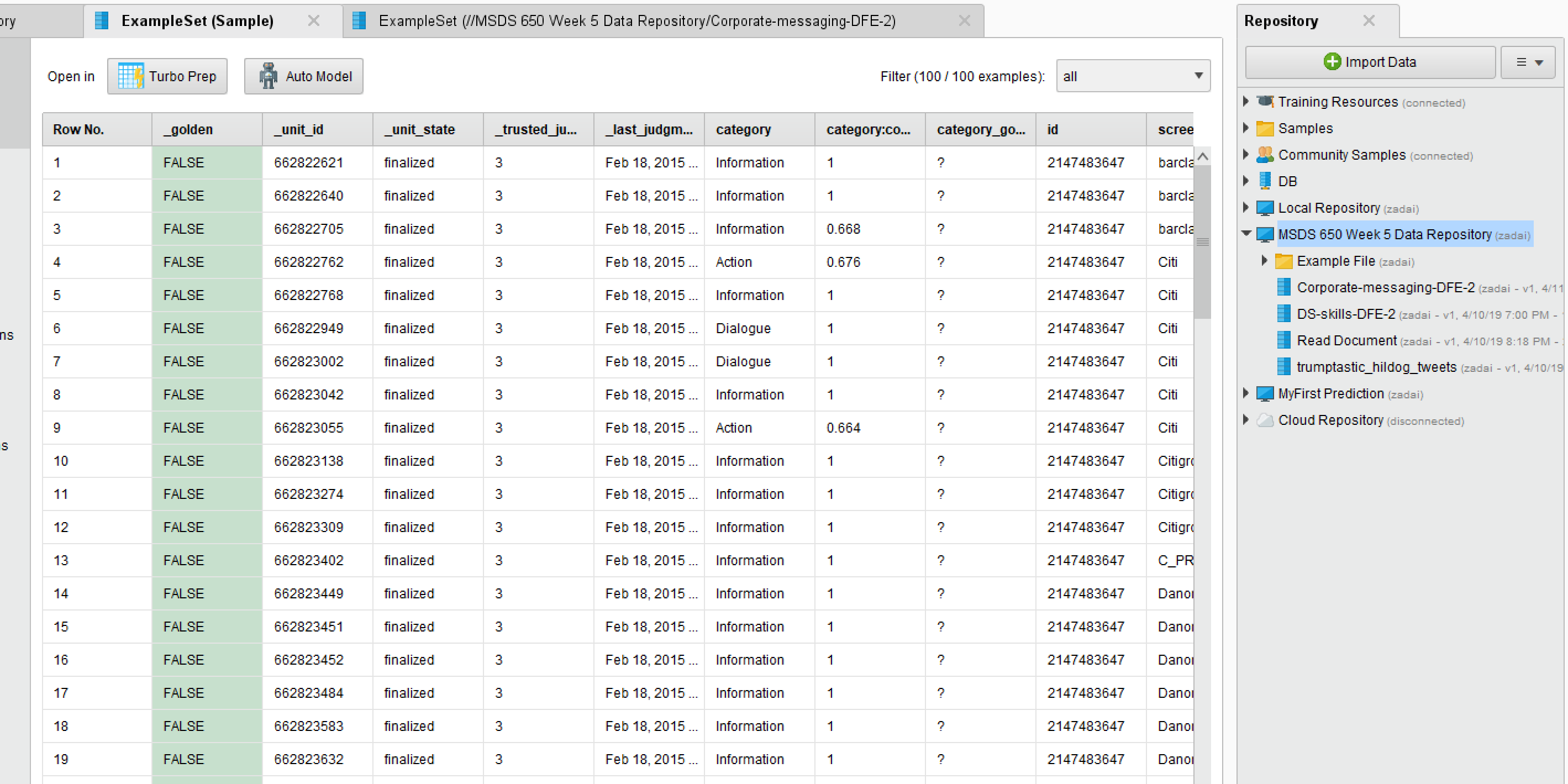
Here I’ve expanded each of the fields and it tells us the field name, type, the number of missing or null values by percentage, a small bar chart, and max and min values.

Now I’m going to look into the 2nd dataset of the assignment, which is the corporate messaging section.

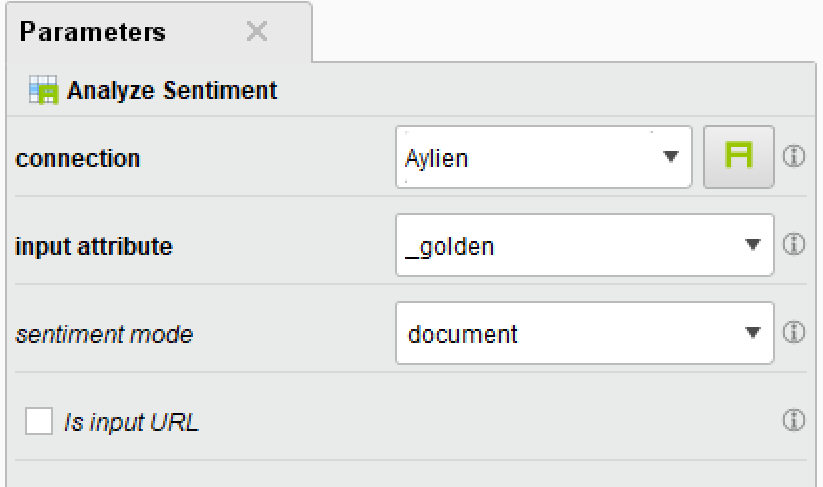
What I first did was, just like the Data Science data, I wanted a look at a sample of the data.



Run the process and now I’m looking at my sample dataset.



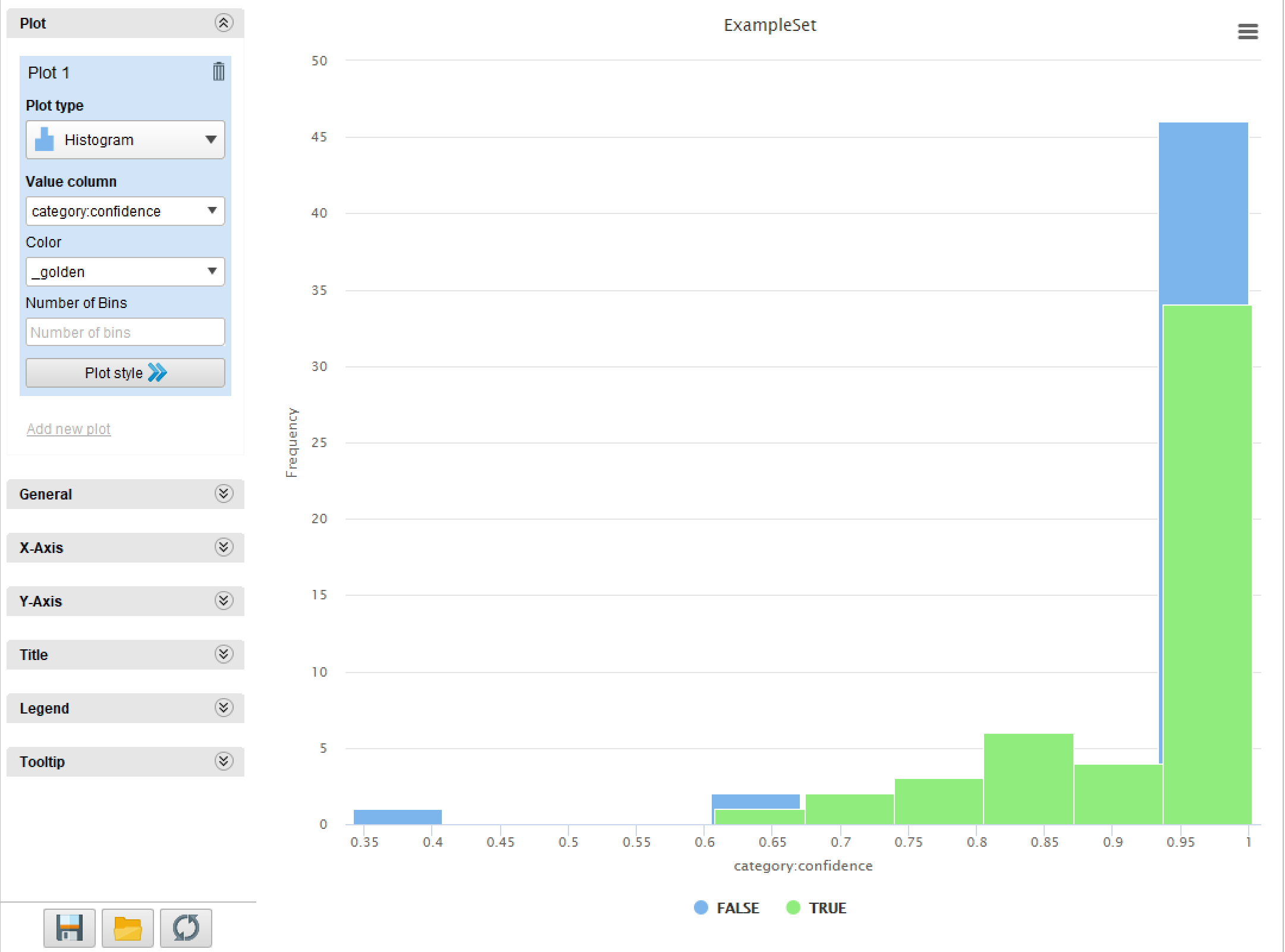
I took my dataset sample and took 50 TRUE and 50 FALSE of the golden field. From here I went in and added the Aylien sentiment operator.



The parameters set that you see above are the input attribute = golden, connection = Aylien, and sentiment mode is set to document.

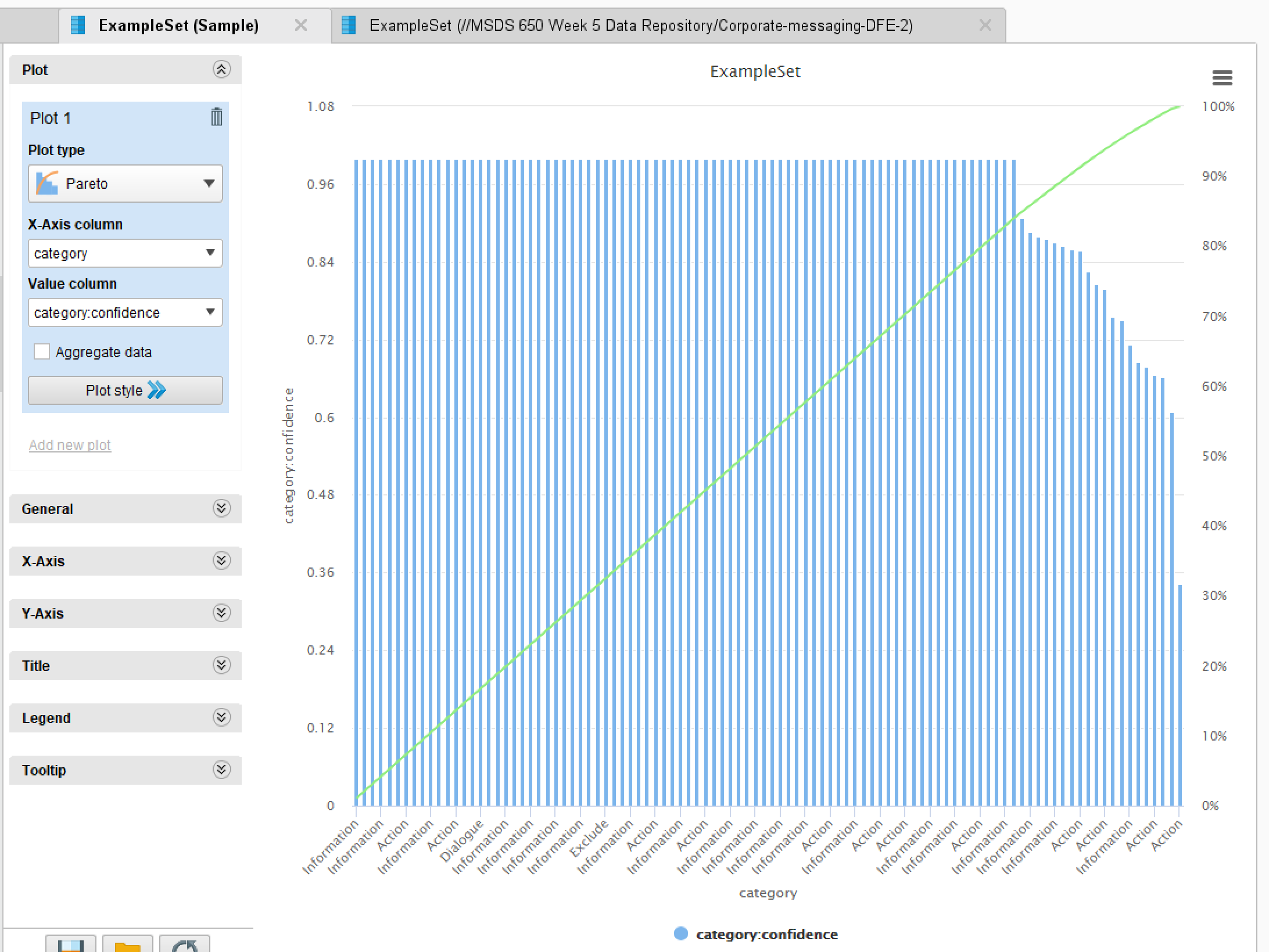
Now I’m going to look at the results and play with some of the visualizations.

The first chart I want to look at is the distribution of category confidence between true and false records.



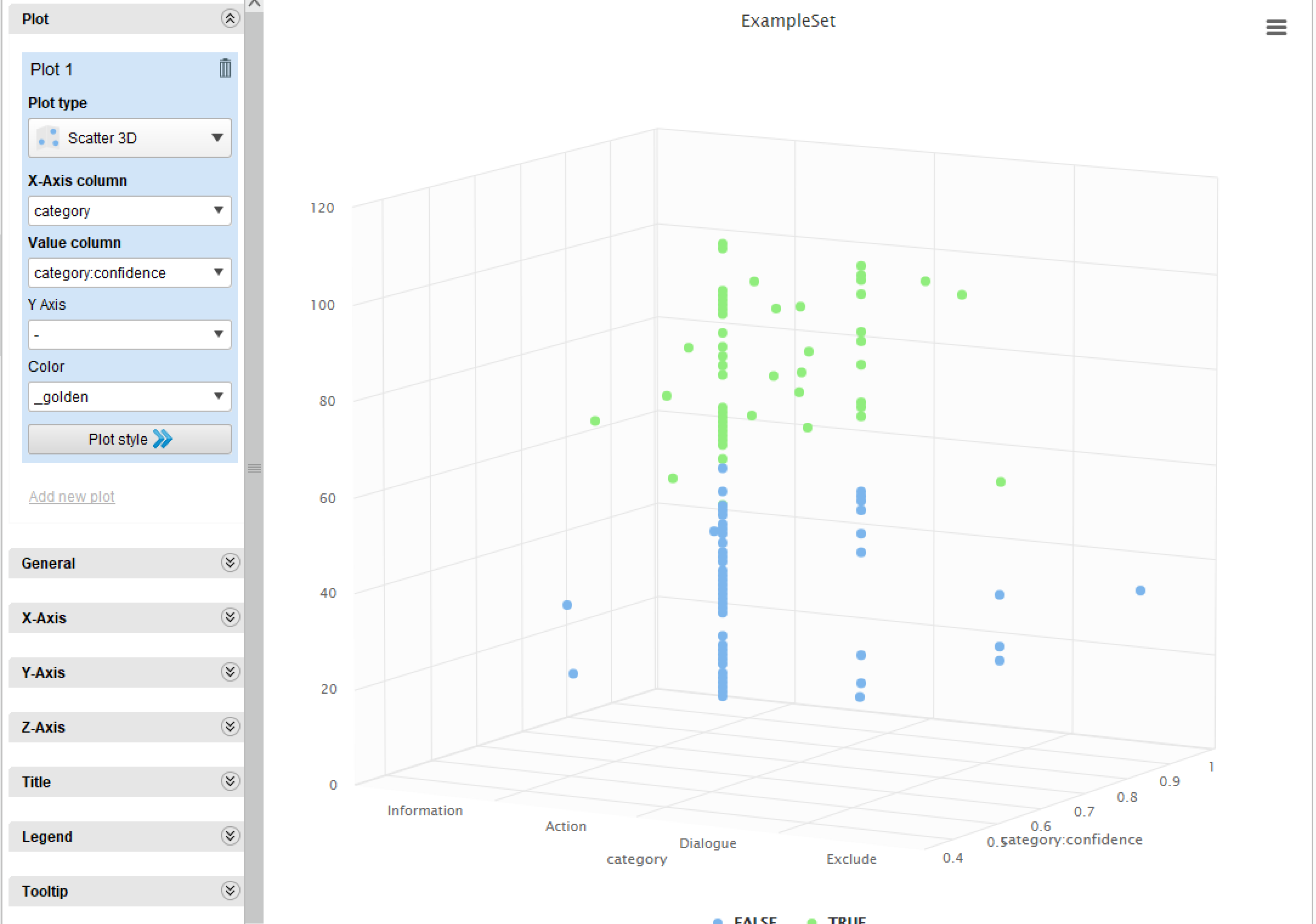
As the chart will show above there was a very low level of low confidence records in our data for either side. The false records had the most records with full confidence with the level at 1. The true values are a little bit more distributed with a little bit of records scattered from 0.6 and 0.95 before a large amount aggregating at 1.

Another chart that I thought was interesting to look at was the Pareto chart, here again I looked at the measure of category confidence which ranges from 0 to 1. With my sample and how my Pareto chart was constructed it took every individual record from the sample and lined them up by their highest levels of confidence to their lowest. The y-axis is set at two different levels which can make someone think the line tallying the total confidence is on the same level as the counted up confidence level but if you look at the y-axis you can see how it is measured based on the side.



From our chart we can see right away that a vast majority of our records have a confidence level of 1, but it tails off for the last 20% of records. Where we see the confidence level start to drop, our totals line begins to tail off and its run isn’t at a 1 to 1 slope like it was before.

The last chart I decided to look at is a 3D scatter plot. The dimensions I wanted to examine in here was the different categories: Information, Action, Dialogue and Exclude; the confidence level and then the number of records total. I also split the color of my records by False and True and that is a color coded difference.



The chart shows that a lot of the records are either Information and Action categories, but for the most part the confidence level regardless of category is extremely high.

From being able to analyze the text from the Corporate messaging data we can distinguish that the confidence level of the records is generally very high and the amount of low confidence is very small. We also know that the categories with the most records in our data are Information and Action, with Dialogue and Exclude being a small part of the records. The distribution of the confidence level really surprised me from what I expected, my expectation was just that the data would be more towards the center but the distribution is high with a drastic tailing off to smaller degrees of confidence.

# Summary

This week’s assignment was a very interesting look at a new application which I have never had the chance to use before this class. Though I feel like my skills are still very raw, I still think it is easy to get data from and be able to sculpt it into meaningful information. I did run into some errors and issues while analyzing the data in Rapid Miner which made it hard for me to continue but when that happened I would eventually pivot and try another method of analysis or I would just use a different set of data. I think as I continue my progression of becoming a Data Scientist I think Rapid Miner can be a great tool to use and create meaningful analysis with.