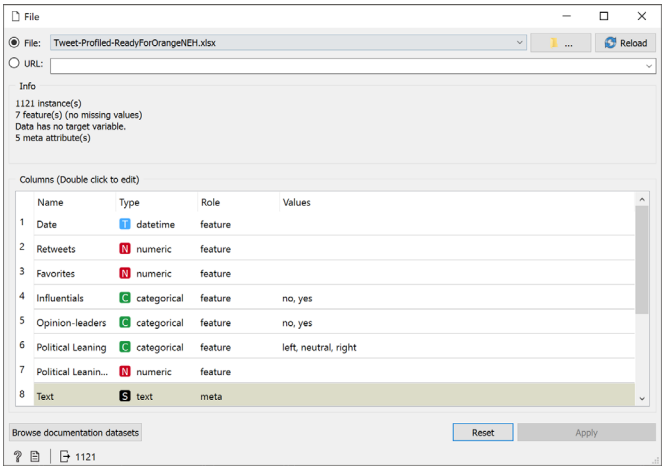
Using the **open folder button**, browse and choose your spreadsheet.

The columns of the spreadsheet will be displayed with the data type, role, and values determined by Orange. In some instances Orange chooses a different type than you want and you can double-click on the type column to change it.

Additionally on this screen it will show you the number of rows (1121 instances), the number of fields it deems are features (e.g. numeric or categorical) and the number of text fields (meta).



File open screen

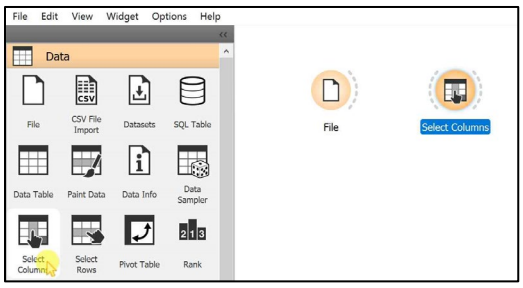
If you change the underlying spreadsheet, you can open this File dialog and click on the Reload button at the top-right to refresh the data.

Once you have selected your file, you can close this dialog.

Use **Ctrl-S** or **File -> Save** to save your workflow. Do this fairly often to ensure you don't lose your work.

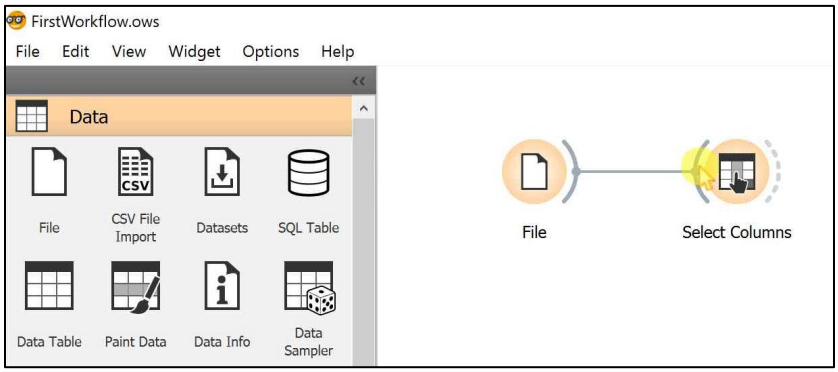
Select Columns

Choose the **Select Columns** widget to add it to your work area. We will use this widget to set up the fields we want to use in our analyses.



Select Columns widget

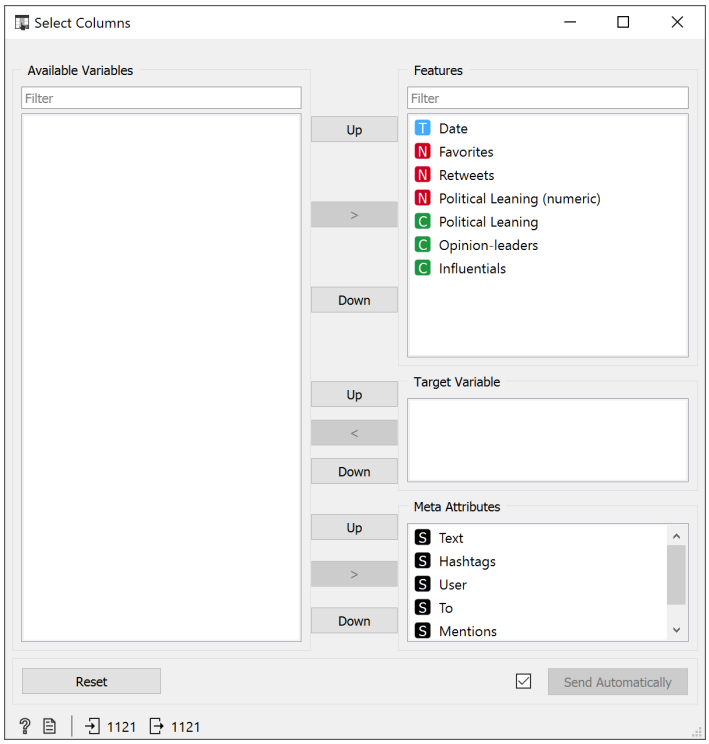
To connect the two widgets together left-click on the output side (right) of the File widget and hold down the mouse button while dragging the connecting line over to the input side (left) of the Select Columns widget.



Connecting the widgets

This connection allows the Select Column widget to see the schema of the spreadsheet file.

Double-click on the **Select Columns** widget to open its options screen.

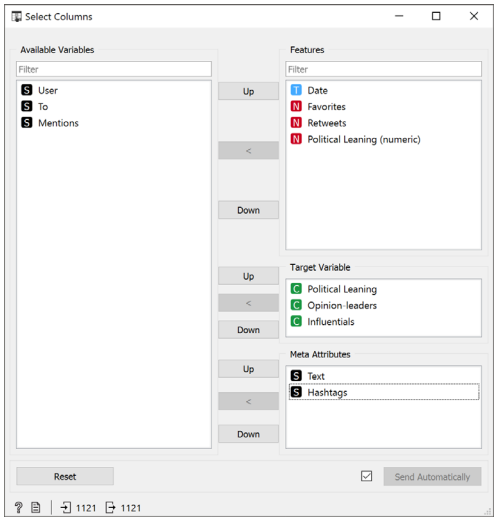


Initial Select Columns screen

Orange will default the spreadsheet fields into the Features and Meta Attributes sections of this screen.

You may **drag and drop** the fields into the different areas of the screen or highlight them and use the arrows in the center section.

For this first workflow, if you are using non-categorized data, leave the target variable area empty. In this tutorial we will not be analyzing the User, To, or Mentions field so those have been moved out of the Meta Attributes area.

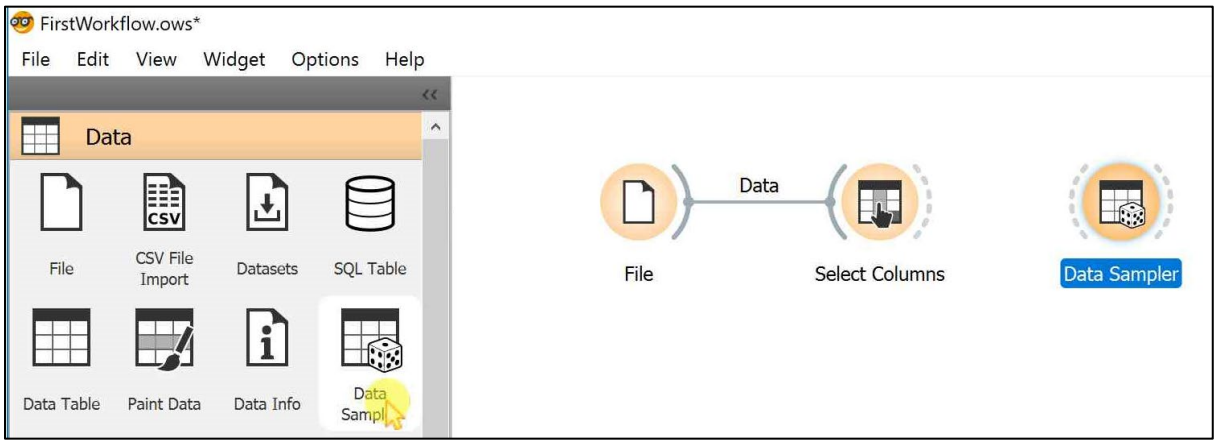


Columns arranged for processing

Once the column names have been arranged in the correct areas, you can close this screen.

Data Sampler

If your dataset is quite large, or just to save time while you develop your workflow, you can use the Data Sampler widget to select a small portion of your data.



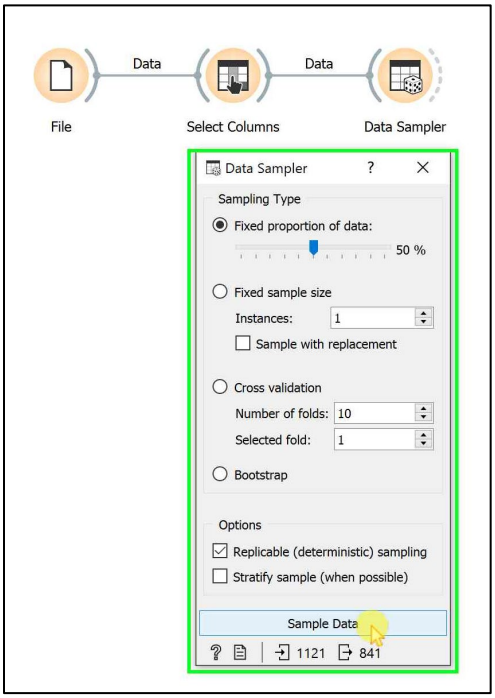
Insert a Data Sampler widget

Connect the output of the Select Columns widget to the input of the Data Sampler widget and then **double-click on the Data Sampler** to open the options screen.

If you have a really large dataset and plan to use a sample throughout your analysis, click on the Data Sampler widget and press F1 to open the help documentation. The documentation for Orange describes each of the options available on a widget and will allow you to make an informed decision on what options to use.

The default for this widget is 75% and if you look at the bottom line in the image below, you will see that the 1,121 rows are reduced to 841. After dragging the slider to 50%, you must click on the Sample Data button at the bottom to update the sample.

For this tutorial, we are going to reduce the size of the sample dataset by half to speed up our initial setup of the workflow. Since you are using your own dataset for this portion of the tutorial, use the **Fixed proportion of data** slider and reduce your data size down to less than 1,000 rows. Once you have sampled your data, you can close this screen.



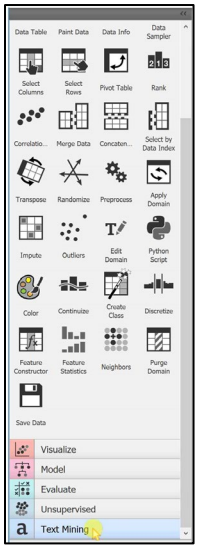
Data Sampler options

Corpus

Most of the text analysis widgets work with a Corpus of texts. Up to this point in our workflow, each of the widgets takes in, and outputs "data." Now that our data has been imported and sampled, we need to convert it to a corpus for further processing.

The Corpus widget is contained in the Text Mining section of the widget toolbox. If you **scroll down** to the bottom of the widget window, you will see various other sections of widgets. We will

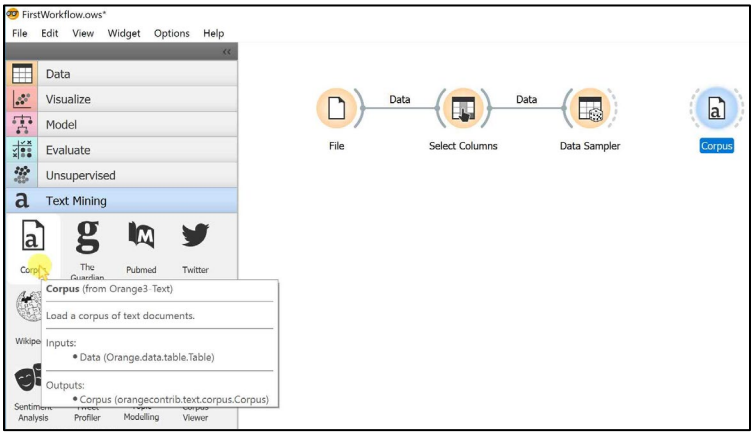
use some of the other sections such as Visualize later but for now, click on **Text Mining** to open that section.



Widget toolbox – **Select Text Mining**

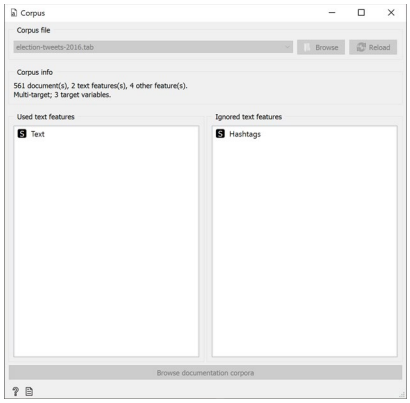
Click on **Corpus** to add the widget to your work area.

When you hover your mouse over a widget in either the toolbox or the work area, it will display a brief indication of the inputs and outputs of the widget. As seen in the image below, the Corpus widget takes Data as its input and sends a Corpus out.



Corpus widget with popup help

Connect the output of the Data Sampler widget to the Corpus widget. Then **double-click** on the Corpus widget to open the options screen.



Corpus widget options screen

This widget screen can be confusing as it is made to be used in two ways. One way is to open a file that is already formatted as a corpus and the other way to use the widget is as we are doing now—converting data to a corpus.

The top section of the screen is greyed out as it is not available when the widget is has input from a data source. The confusion is that they show a default Corpus file name that is not our input file name. The other confusion could appear if you open this widget options screen before you connect the Data Sampler widget. If that happened, you would see different fields listed in the two columns of features. So be sure to connect this to your Data Sampler first and ignore the filename at the top.

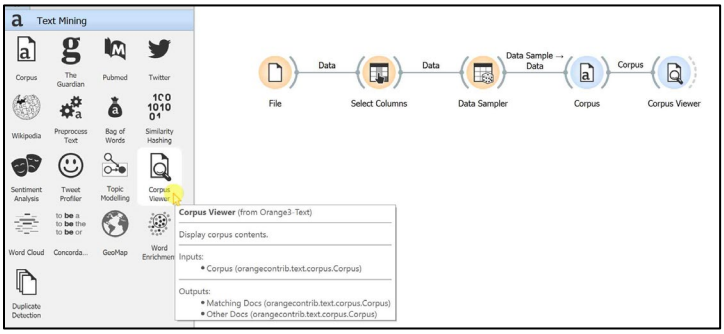
Drag the Hashtags field from the Ignored text features to the Used text features. Once both fields are in the left column, close this window.

Corpus Viewer

The Corpus View widget allows you to look at your data within Orange.

Add the widget to your work area and connect it to the Corpus widget.

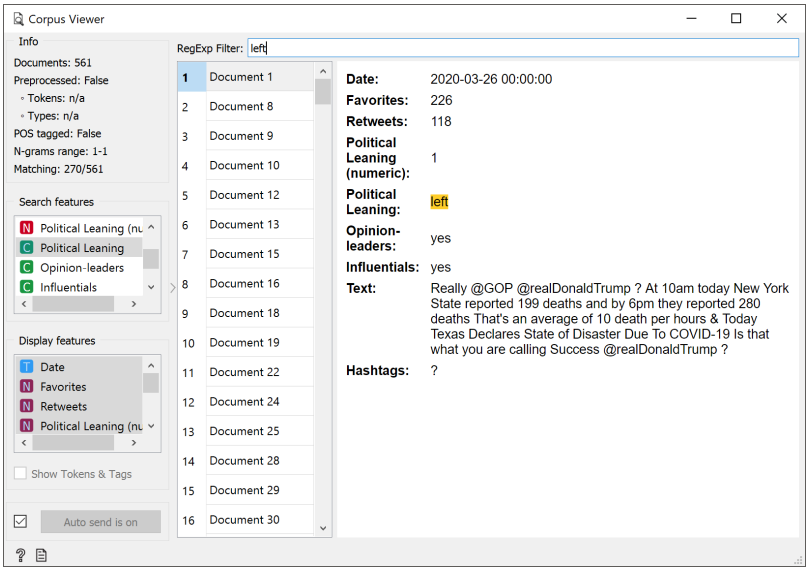
As shown in the image below, the line between the Corpus and Corpus Viewer shows the new data type as Corpus.



Corpus Viewer widget in work area

Double-click the Corpus Viewer widget to open the options screen. You will need to drag the scroll bar between the list of Documents and the right-hand section of the screen to make it legible.

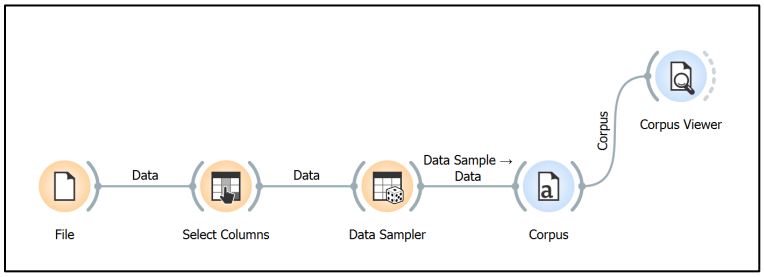
This window has three columns. The first shows you some of the metadata such as the number of documents in the set (our 50% sample) and the features you can use to Search and filter the documents to view. In the image below, Political Leaning was selected as the search field and "left" entered in the top RegExp Filter field. The documents now listed are the tweets categorized as having users that are left leaning.



Corpus Viewer widget options screen – **Select** a **Search feature** in the left panel and then at the top in the RegExp Filter field, enter a term to use.

The checkbox and button at the bottom left (Auto send is on) is similar to the Automatically Commit buttons we have seen in earlier widgets. If this viewer/filter is connected to another widget, it would automatically send the data on to the next widget if that is checked.

Although the widgets default to being added in a row across the screen, you are able to drag and drop them anywhere in the work area. In the image below, the Corpus Viewer has been moved above the general line of the workflow.



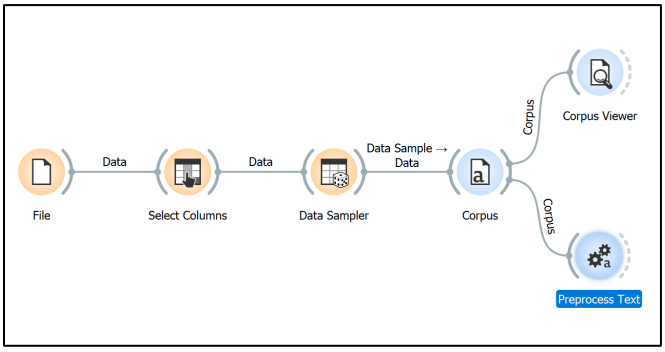
Current workflow

Your workflow should now include the widgets shown above. We have imported our data file; chosen the columns for features, target variables, and meta variables; sampled our dataset down to less than 1,000 rows; turned that smaller dataset into a corpus; and viewed and filtered the corpus. We are now ready for text processing.

Preprocess Text

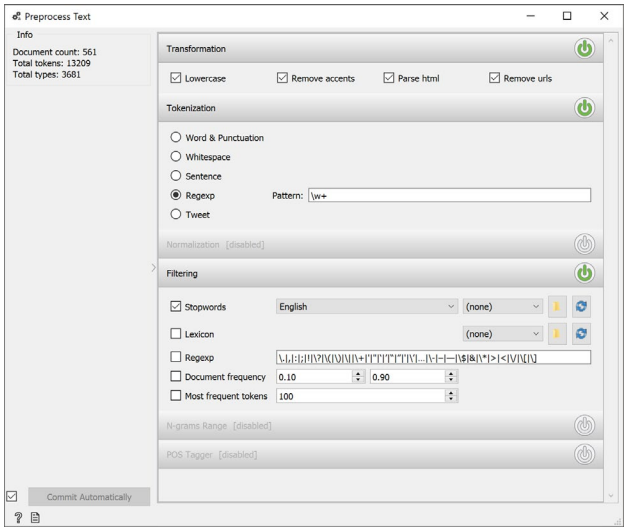
The output from a widget can be sent to multiple widgets. This saves us from having to create separate workflows that duplicate steps.

Add a Preprocess Text widget to your work area like shown in the image below.



One-to-many widgets

Double-click the Preprocess Text widget to open the options screen. The green on/off buttons activate sections of this screen. By default, the Transformation, Tokenization, and Filtering options are on. The preprocessing splits the text (in our case tweets) into tokens that are used in further analysis.



Preprocess Text widget options screen

By checking the boxes in the Transformations section it allows you to remove HTML tags, URLs, and accents from characters (i.e. it uses **e** instead of **è**). The Tokenization section is set to use regular expressions to break the tweets apart by word and the Filtering section has Stopwords checked and the language set to English. If your tweets are in a language other than English, you should switch the language.

Stopwords are common words such as "an," "the," "than," etc. Since these words are very high frequency in most texts, they are filtered out from the tokenized words so that they do not skew or clutter up any of the results.

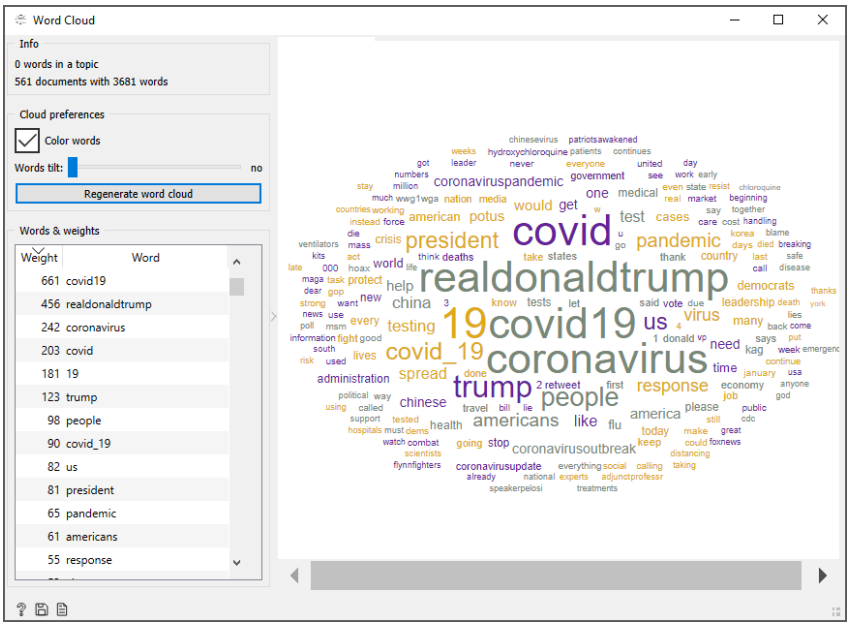
Note the Commit Automatically checkbox at the bottom left.

Make your selections match the sample screen above and close this window.

Word Cloud

The last widget we will add to this first workflow is the Word Cloud.

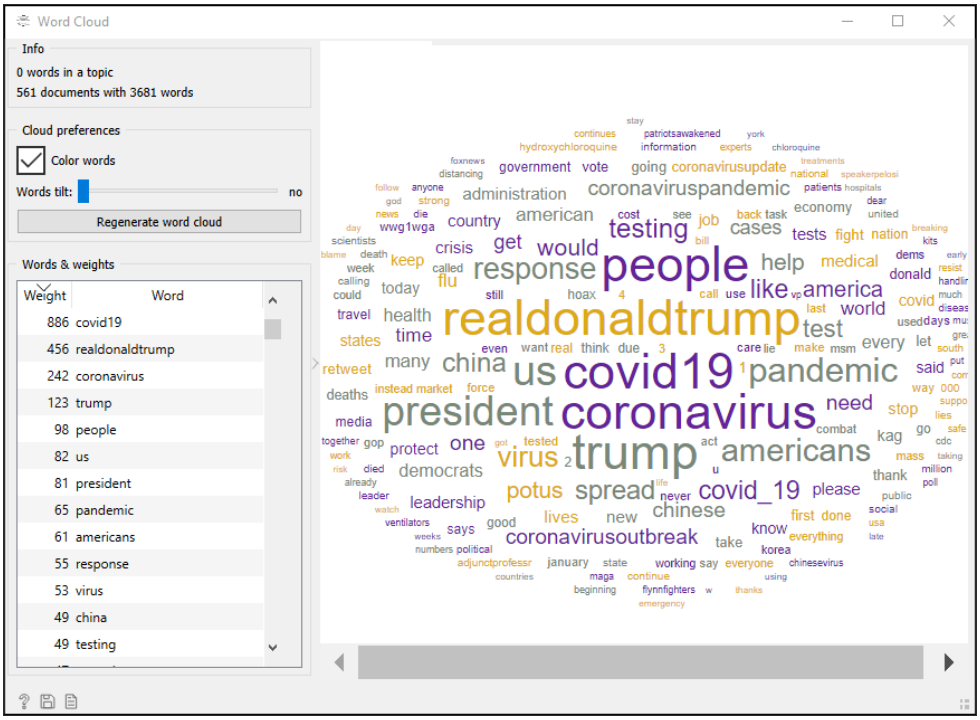
Add the widget to your work area, **connect** it to the Preprocess Text widget and **double-click** on the Word Cloud to open the options screen.



Word Cloud widget options screen

On the left side is a list of the words contained in the cloud and the number of occurrences of each. Remember that this is currently a smaller sample of the data. In looking at the fourth and fifth items ("covid" and "19") we can see that the tokenization routine split "covid-19" on the hyphen and treats them separately. Also notice that covid_19 (with an underscore) is also treated separately.

By going back to the spreadsheet data and replacing "covid_19" and "covid-19" with "covid19" in the tweet text and then reloading the data in the File widget, the Word Cloud automatically updates to the image below.



Word Cloud from the sampled data

You may note that covid_19 still exists in the word cloud even though the tweet text was updated in the spreadsheet. This is because we included the hashtags back when we created the Corpus widget. If you want to just focus on the tweets versus the tweets and hashtags, you could go back and remove the hashtags from the Corpus widget's "Used features" list.

To download a copy of your word cloud (or any other Orange visualization), use the Save icon at the bottom left of the window. This option allows you to save it as a PNG, SVG, or PDF.

Tutorial Deliverables – First Workflow

For attendees of the NEH Digital Culture Institute, please submit the following as a PDF document:

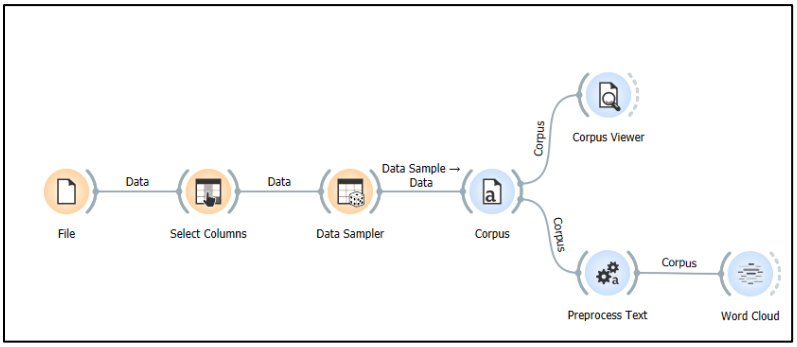
1. A short paragraph describing how you might categorize the tweet data you have for your hashtag topic
2. A screen shot of your completed workflow
3. A PNG of your word cloud

This completes the "First Workflow -> File Load to Word Cloud" section of the tutorial

Second Workflow -> Preprocess Text to Sentiment Analysis

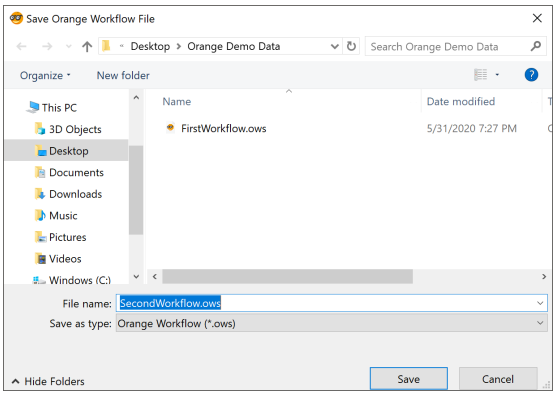
We will be using the **COVID-19 sample dataset** for this tutorial since it already has been categorized. We will also use the first tutorial's workflow as a base to shorten the number of steps in this second workflow.

In Orange, **click** on **File -> Open and Freeze** and open your saved workflow from the first tutorial.



Workflow from first tutorial

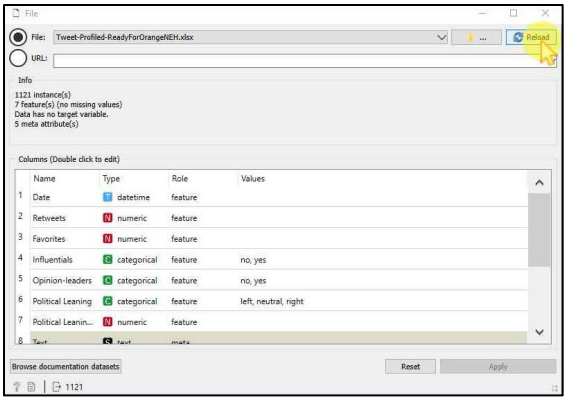
Then **click** on **File –> Save As** and name it **SecondWorkflow.ows**.



Save As to start the second tutorial

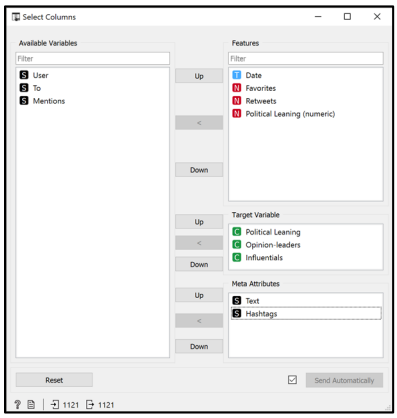
Navigate to http://chdr.cah.ucf.edu/neh-digculture/Tweet-Profiled-ReadyForOrangeNEH.xlsx and download the Tweet-Profiled-ReadyForOrangeNEH.xlsx spreadsheet.

In your workflow, **double-click** the **File** widget and browse to where you saved the spreadsheet. Remember if you saved it on your Desktop you may need to navigate through your Users directory to get to the file.



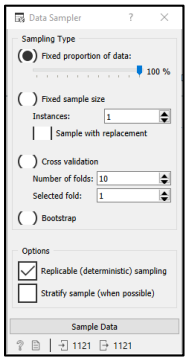
Reload the new data file – **Select** the new spreadsheet and **click on Reload**. When the number at the bottom left reads 1121, the file has been loaded and you can close this window.

Double-click on the **Select Columns** widget.



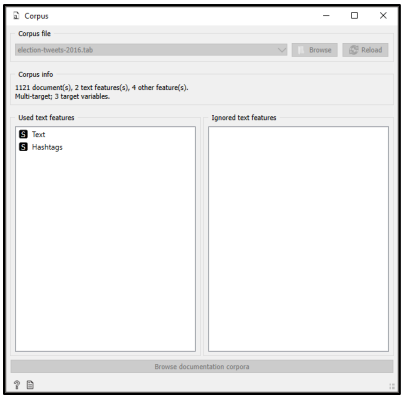
Select columns widget – **Arrange the columns to match this screen shot**. When it matches, you can close this window.

Double-click on the **Data Sampler** widget and set it to **100%** as we want to use the entire dataset. Remember to click on the Sample Data button at the bottom after changing the slider.



Data Sampler widget – **Set slider to 100% and click on Sample Data**. Close window when complete.

Double-click on the **Corpus** widget and make sure both Text and Hashtags are in the "Used text features" column.



Corpus widget – **Put Text and Hashtags in left column**. Close window when complete.

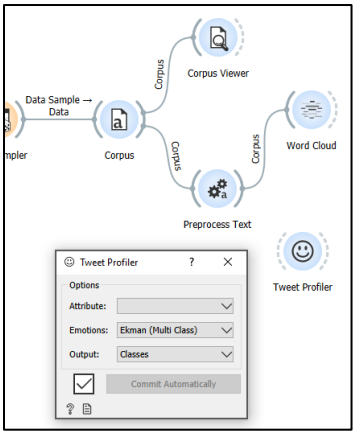
We will leave the **Preprocess Text** widget alone unless you had changed the Stopwords language in the previous tutorial. If you did, please set it back to English as the tweets in this dataset were constrained to be in English.

Tweet Profiler

The Tweet Profiler widget takes a fair amount of time to run on this tutorial's dataset.

Add the **Tweet Profiler** widget to the work area, but do **NOT** connect it to any other widget yet.

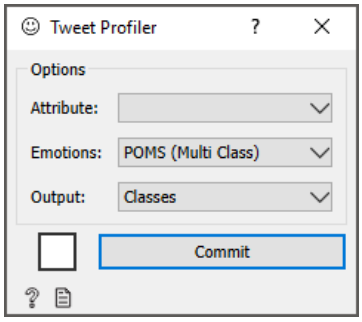
Double-click on the widget to bring up the options screen. The defaults are shown in the figure below.



Tweet Profiler widget – Default configuration

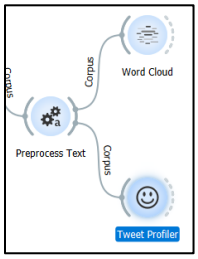
The article "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model" by Niko Colnerič and Janez Demsar describes in detail the various models used in the Tweet Profiler widget. If you would like more information on how the models are derived.

Choose "POMS (Multi Class)" for the Emotions model and **uncheck** the box next to Commit. This will prevent model from running immediately once it is connected to the workflow.



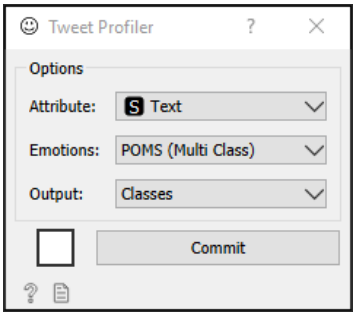
Tweet Profiler widget – **Select POMS (Multi Class) and uncheck Commit**. Leave this window open for now.

Connect the Preprocess Text widget to the Tweet Profiler.



Preprocess Text to Tweet Profiler

Select the **Tweet Profiler** window again. The Attribute should now be set to Text (the tweet content).

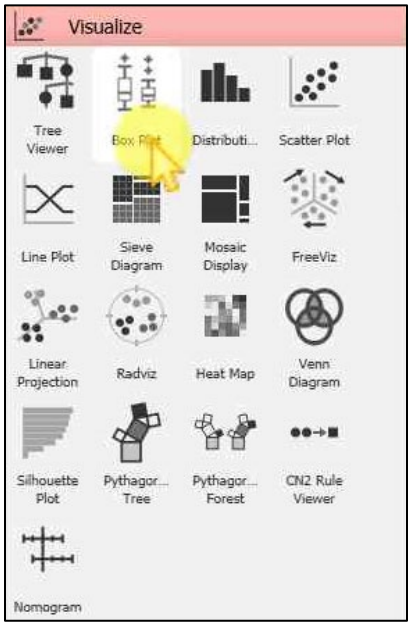


Tweet Profiler with Attribute set to Text

Click on the **Commit** button and be patient as it runs. The percent complete should show in the bottom under the Commit button.

When the Tweet Profiler is complete you can close the window.

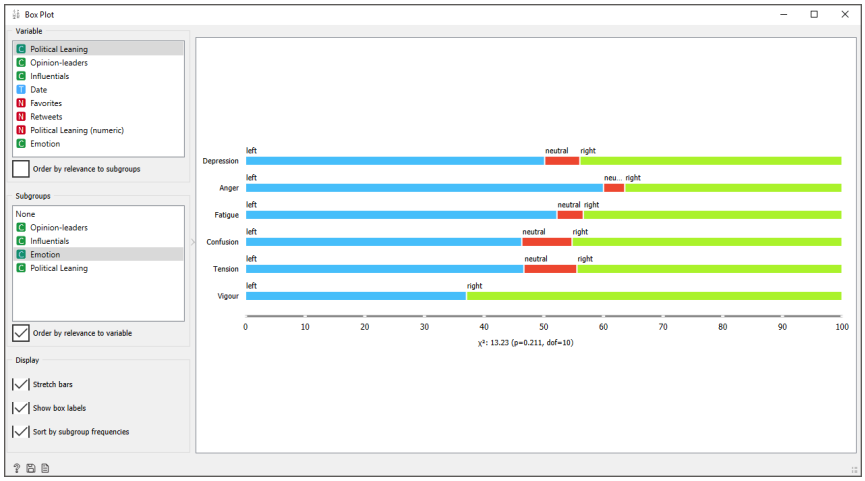
Select the **Visualize** section of the Widget toolbox and **choose Box Plot**.



Visualize widgets – **Select Box Plot**

Connect the **Tweet Profiler** widget to the **Box Plot** widget and then **double-click** on the Box Plot.

The column on the left contains the variables that you can investigate through the box plot. The image below shows the emotions subgroups in comparison to the political leaning category.

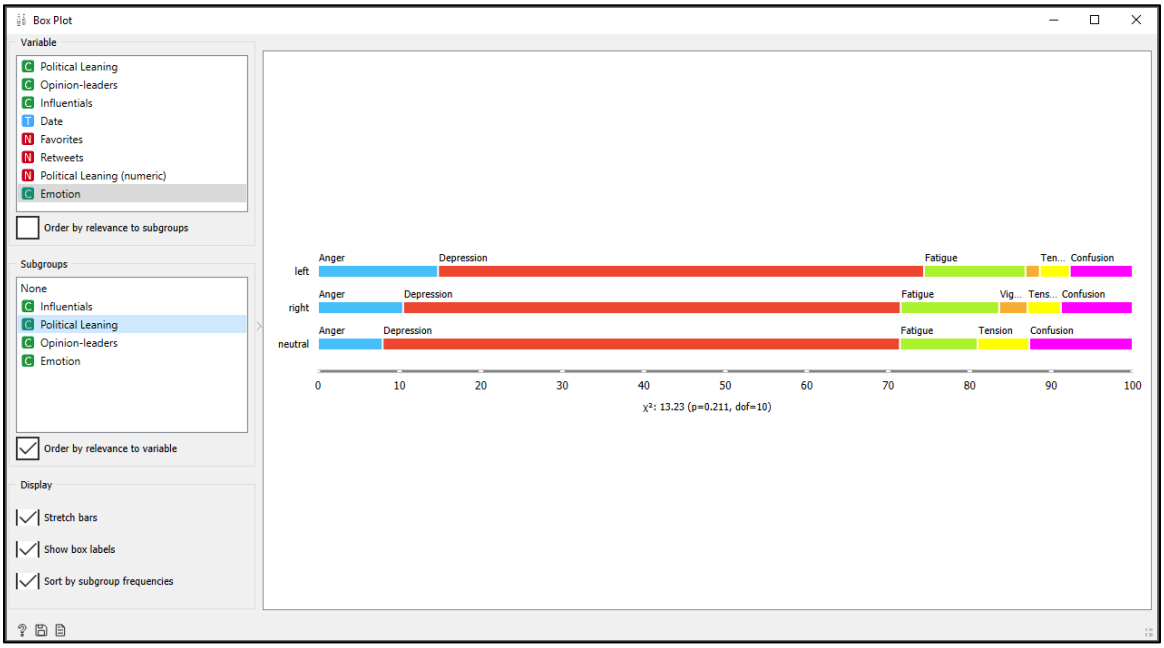


Box Plot of POMS Multi Class – Emotion subgroups by Political Leaning

The POMS (Profile of Mood States) multi-class model scores emotions in 6 categories (Depression, Anger, Fatigue, Confusion, Tension, and Vigour). In the article by Colnerič and Demsar, they used specific adjectives in training their model.

If you hover your mouse over the sections of each of the plot's lines, it shows you the values. For example, left politically leaning is 50.15%, neutral 5.9%, and right 43.95% on the plot for depression. Just visually scanning this plot, the left is more angry, tired, and depressed, and the right is much higher (62.96%) on the vigor line. The adjectives used for vigor include: active, energetic, full of pep, lively, vigorous, cheerful, carefree, and alert.[3]

If you flip it and look at Political leaning subgroups by emotions, you can see that those who are neutral politically (in the narrow view defined by our dataset), their tweets show as more depressed and confused.



Box plot for Political subgroups by Emotions

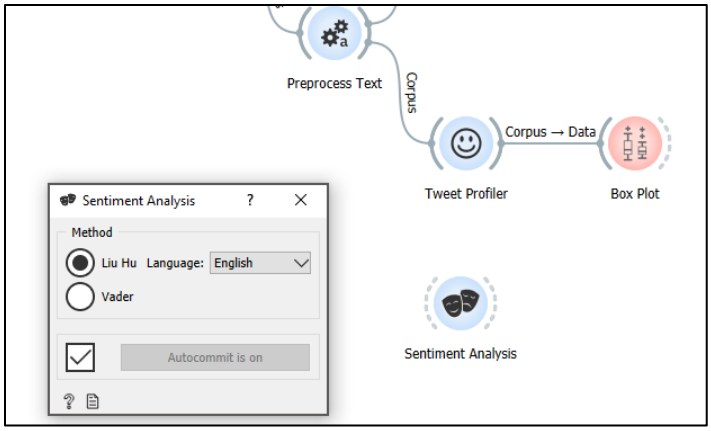
After examining some other combinations, please close the box plot window.

Sentiment Analysis

For the last section of the tutorial, we are going to look at two other methods of sentiment analysis.

Back in the **Text Mining** portion of the Widget toolbox, add **Sentiment Analysis** to the work area and **double-click** to open it.

[3] Colnerič and Demsar, p. 2.



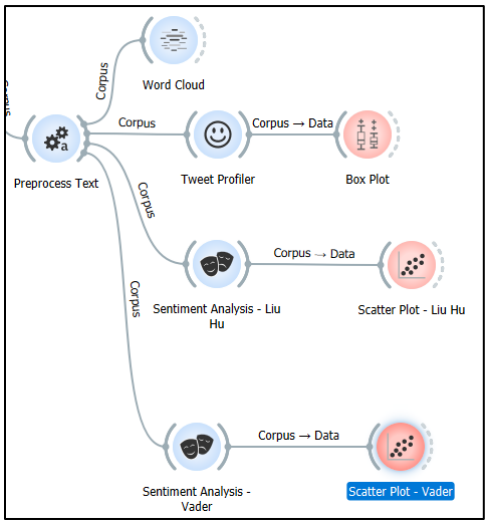
Sentiment Analysis widget options

The Sentiment Analysis widget has two options for the method. Liu Hu is specifically lexicon based and language specific and ranks sentiment as a single number that is either negative, zero, or positive. Vader is uses rules-based analysis in addition to the lexicon and has 4 scores: positive, negative, neutral, and compound.

Rename the Sentiment Analysis widget to "Sentiment Analysis – Liu Hu" and make sure Liu Hu is the selected method.

Add a second widget, set the Method to **Vader**, and **rename** the widget to "Sentiment Analysis – Vader"

From the Visualize widget menu, **add two Scatter Plots** and **rename** them according to the sentiment methods. **Connect** each pair to the **Preprocess Text**. The end of your workflow should look similar to the image below.

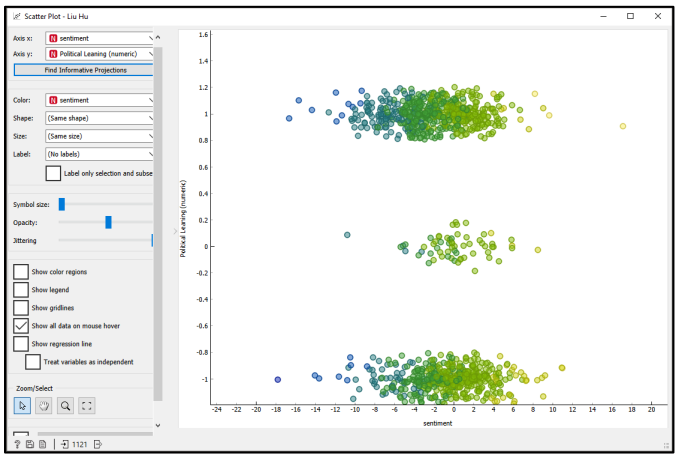


Two Sentiment Analysis and Scatter Plot pairs

Double-click on the **Liu Hu Scatter plot**. For the **x-axis**, select **sentiment** and for the **y-axis** select **Political Leaning (numeric)**. In an earlier version of Orange, they allowed categorical variables to be used in scatter plots. Since they removed that option, the political leanings category (right, neutral, left) was copied to another column in the spreadsheet and the values set to (right -1, neutral 0, left 1) so that it would be available for evaluation.

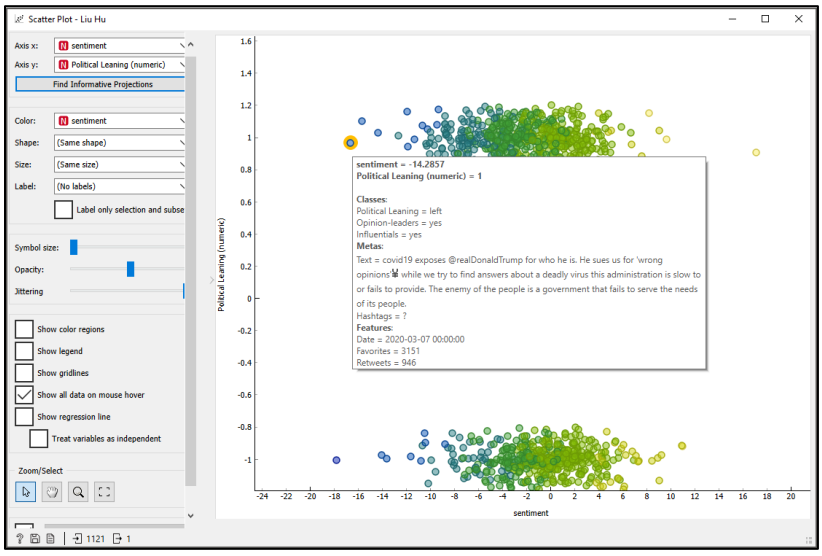
Select sentiment for the Color, **adjust** the Symbol size to the far left (small) and the Jittering to the far right (wide). This spreads the data out to see the pattern easier.

Negative sentiment will be to the left and positive to the right.



Liu Hu Scatterplot of Sentiment and Political leaning

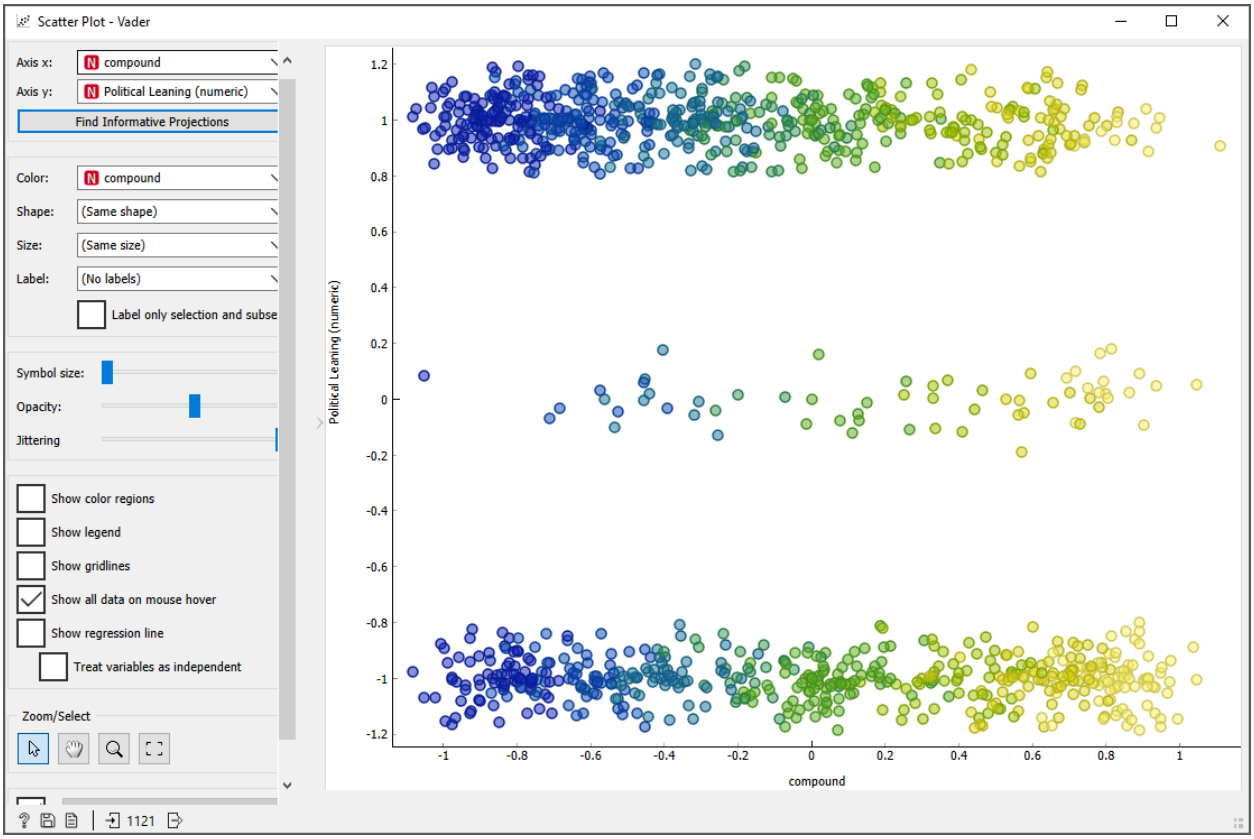
If you click on specific nodes within the scatter plot to select them and then hover over the selected node, a box will be displayed with the tweet information.



Most negative left-leaning tweet – Liu Hu

Double-click on the Vader scatter plot widget. **Click** on the x-axis dropdown and note that it includes the 4 scores from this model. **Select** Compound for the x-axis as this is similar to what the Liu Hu model has (a single value).

Select Compound for Color and **set** the symbol size to small and the jittering to high.



Vader Scatterplot of compound sentiment and political leaning

Note that the density of the left political leaning plot (at the top) is denser on the left side (negative) and the right political leaning plot (at the bottom) is denser on the right side (positive). This correlates with the Tweet Profiler where it showed left political leaning as more angry and the right as having more vigor (cheerful, carefree, etc.)

Depending upon your research questions and how you categorize your own data, sentiment analysis may not be a helpful tool for you. But since Twitter and most social media show emotional charged content, these tools will be a good place to start your analysis.

Tutorial Deliverables – Second Workflow

For attendees of the NEH Digital Culture Institute, please submit the following as a PDF document:

1. A screen shot of your completed workflow

2. A PNG of your Tweet Profiler Box Plot (your choice of variables)
3. A PNG of one of the Sentiment Analysis Scatter Plots (your choice of method and varoables)

This completes the "Preprocess Text to Sentiment Analysis" section of the tutorial

Concluding Thoughts

By following the methods outlined in this tutorial, you will be able to do basic text analysis, including sentiment analysis, on a dataset of tweets.

Orange is a powerful data mining tool that has many features for different data types. We have just touched on a few of the text analysis and visualization widgets. The workflow concept of connecting widgets allows you to analyze data without doing any programming. This makes the analysis more accessible to those with early computing and/or statistical knowledge.

The takeaways from this tutorial are to remember the following:

1. Spend your effort on cleaning and exploring your data before you ever bring it into Orange
2. Remember that each of the widgets will process automatically unless you tell them otherwise and this, along with the size of your dataset, can sometimes lead to a "not responding" message even though the program is processing.
3. Take some time to watch the other Orange channel tutorials on YouTube as they may give you other ideas on ways to analyze a variety of data: https://www.youtube.com/channel/UClKKWBe2SCAEyv7ZNGhIe4g/playlists

If you have any questions, I can be reached at AmyG@ucf.edu.