Orange Tweet Analysis Tutorial

Abstract

This tutorial walks you through analyzing Twitter data using the data mining tool Orange. We will use a few of the text analysis tools to learn the basics. The tutorial requires a dataset of tweets in an Excel spreadsheet. To collect your own, please see the Twitter Data Scraping Tutorial for additional information. A sample dataset will be supplied.

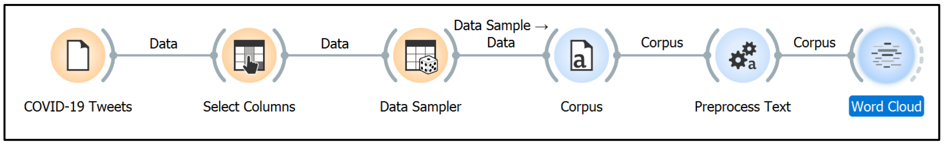
Amy Larner Giroux, PhD
AmyG@ucf.edu

Table of Contents

Introduction

This Orange Tweet Analysis tutorial will step you through the process of installing Orange, pre-processing your tweet data, importing the tweet data from a spreadsheet, and analyzing the textual content of the tweets in various ways.

Orange is a data mining toolset and uses a workflow process enacted through what they call "widgets." The widgets perform various functions from reading data files to creating word clouds. The image below shows a sample workflow.



Orange workflow example

Some assumptions were made when designing this tutorial to allow it to be as comprehensive as possible for a very broad audience. The explanations are written for users who:

1. do not have Orange installed
2. are familiar with Twitter and the concept of hashtags and usernames
3. have a raw dataset of tweets in Excel format
4. have minimal experience with text analysis

If you are already familiar with parts of these concepts, or have completed steps such as installing Orange, please just skim the instructions where you feel confident with the process.

There are a set of outcomes for this Orange data analysis tutorial and the instructional material is separated into these goals:

1. Install Orange and the Text Add-on
2. Pre-processing your data before using Orange
3. Learning some overall Orange concepts/tips
4. First workflow -> file load to word cloud
5. Second workflow -> preprocess text to sentiment analysis

Some of the steps to follow are embedded in the main text of the instructional material, while others are in the captions of the figures. **Bolded text** has been used to draw your attention to items to do. In some of the illustrations, **a yellow cursor** is visible to indicate what to select.

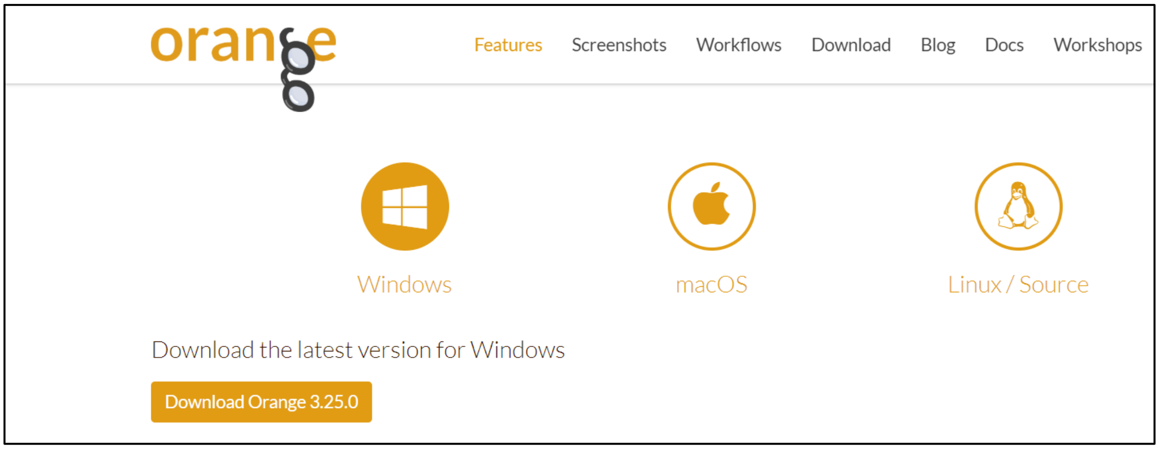
NEH Digital Culture Institute Deliverables are listed at the end of the First and Second workflow sections.

Install Orange and the Text Add-on

Orange is an open-source data mining and analysis tool that uses widgets to create workflows to process the data. We will be using the basic functionality of the program along with the Text add-on. Orange is available for Windows, Mac, and Linux, and the installation is straightforward.

Navigate to: https://orange.biolab.si/download

At the top of the page are the links to the installers for the various operating systems.

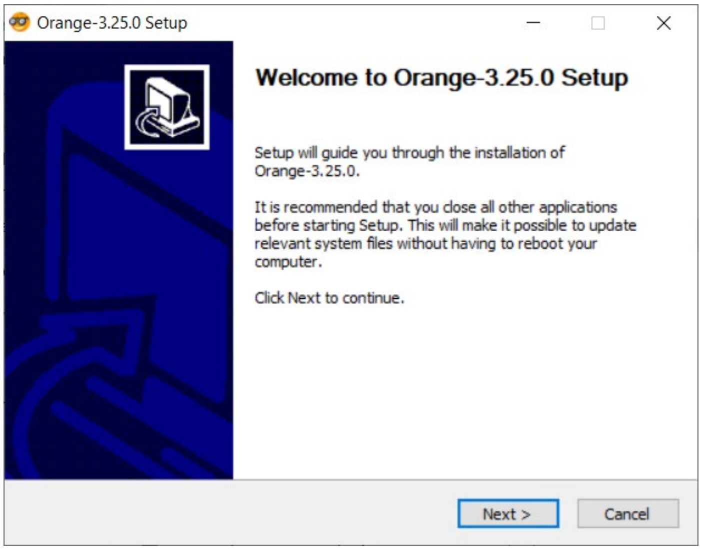


Download the latest version for Windows

Orange Installer Options

Select the appropriate installer for your operating system to save the installer to your computer.

Launch the installer. The installation wizard will move you through the various steps and allow you to adjust a few options. The sequence of screens is shown below.



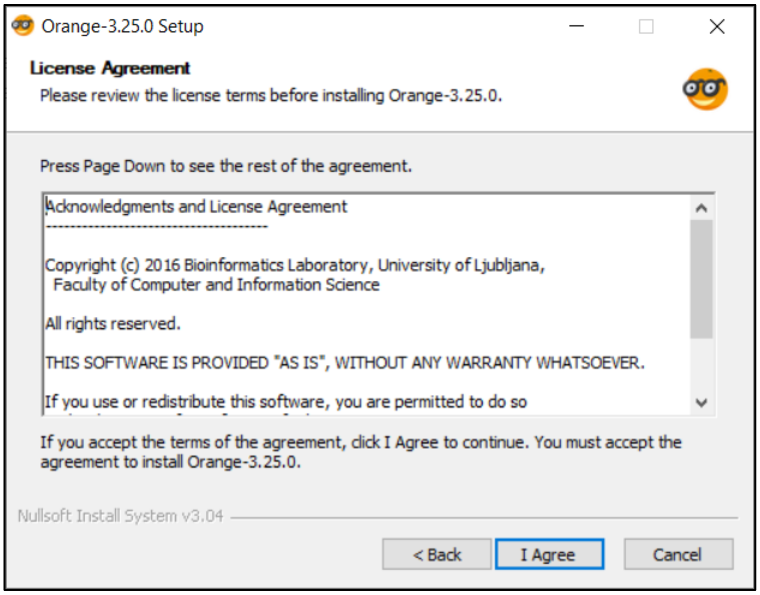
Welcome to Orange-3.25.0 Setup

Setup will guide you through the installation of Orange-3.25.0.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Welcome Screen – Click **Next**



Orange-3.25.0 Setup

License Agreement

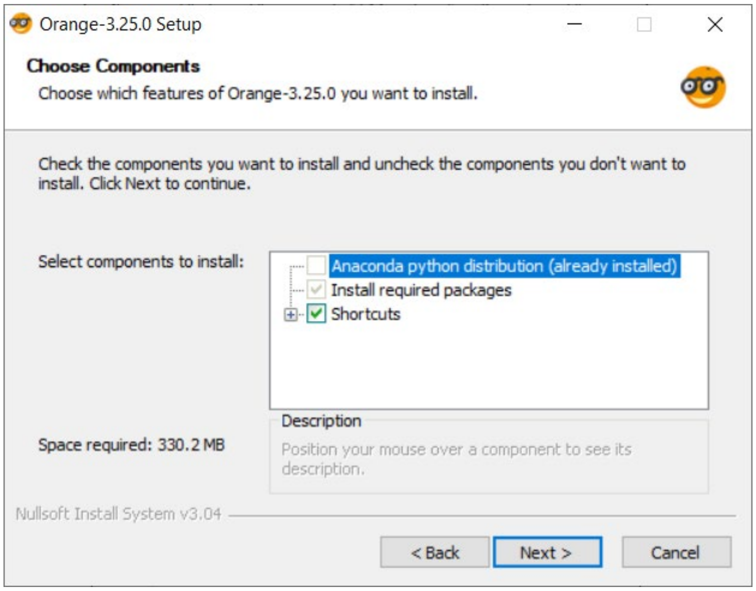
Please review the license terms before installing Orange-3.25.0.

Press Page Down to see the rest of the agreement.

```
Acknowledgments and License Agreement
-------------------------------------

Copyright (c) 2016 Bioinformatics Laboratory, University of Ljubljana,
 Faculty of Computer and Information Science

All rights reserved.

THIS SOFTWARE IS PROVIDED "AS IS", WITHOUT ANY WARRANTY WHATSOEVER.

If you use or redistribute this software, you are permitted to do so
```

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Orange-3.25.0.

Licensing Screen – Click **I Agree**



Orange-3.25.0 Setup

Choose Components

Choose which features of Orange-3.25.0 you want to install.

Check the components you want to install and uncheck the components you don't want to install. Click Next to continue.

Select components to install:

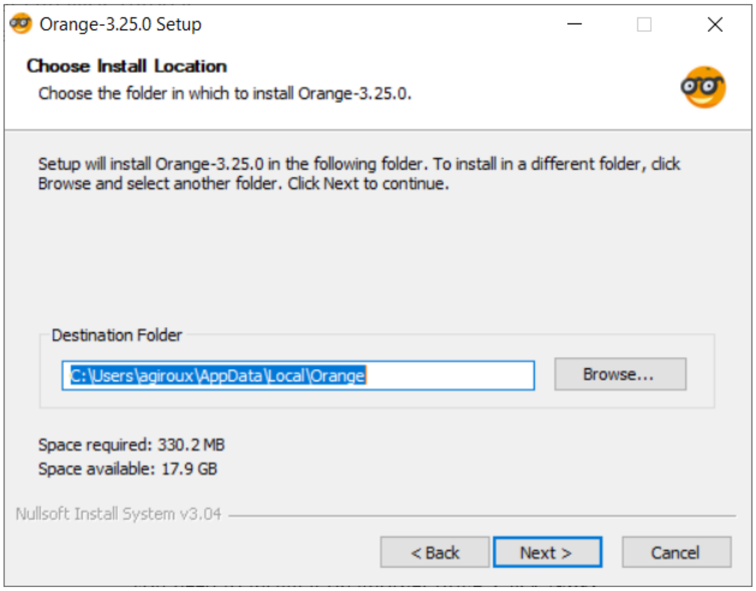
- □ Anaconda python distribution (already installed)
- ☑ Install required packages
- ☑ Shortcuts

Description

Position your mouse over a component to see its description.

Space required: 330.2 MB

Choose Components – If your computer does not have Anaconda installed, that checkbox will be selected and the installation will include those components. Click **Next**



Orange-3.25.0 Setup

Choose Install Location

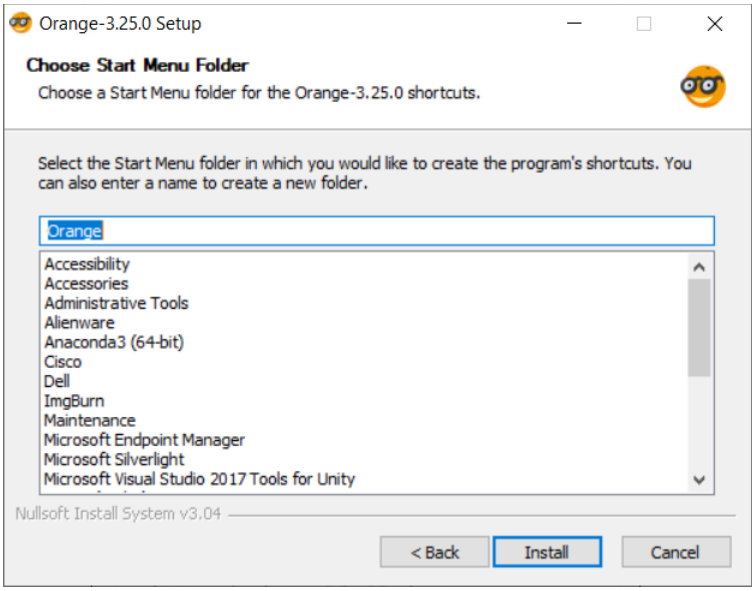
Choose the folder in which to install Orange-3.25.0.

Setup will install Orange-3.25.0 in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder: C:\Users\agiroux\AppData\Local\Orange

Space required: 330.2 MB
Space available: 17.9 GB

Installation Location – Typically leave the default unless you need to install it on another drive. Click **Next**



Orange-3.25.0 Setup

Choose Start Menu Folder

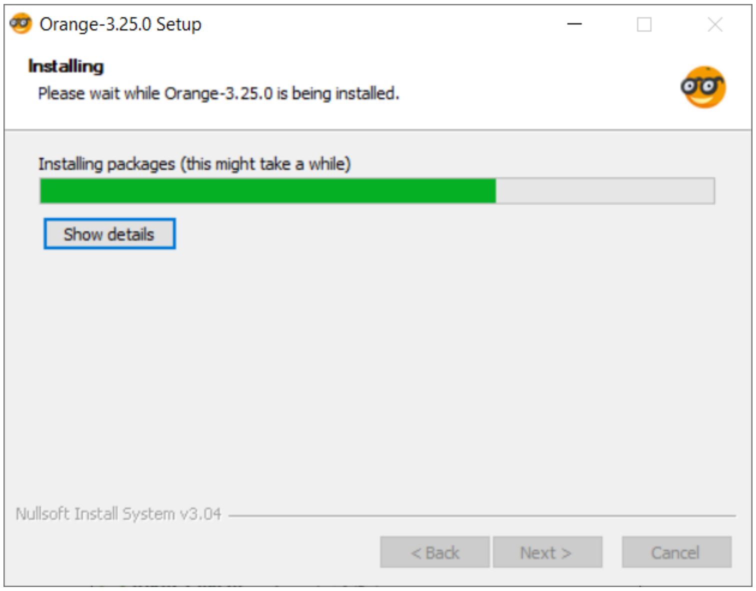
Choose a Start Menu folder for the Orange-3.25.0 shortcuts.

Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.

Orange

Accessibility
Accessories
Administrative Tools
Alienware
Anaconda3 (64-bit)
Cisco
Dell
ImgBurn
Maintenance
Microsoft Endpoint Manager
Microsoft Silverlight
Microsoft Visual Studio 2017 Tools for Unity

Start Menu – leave these options as is and click **Install**

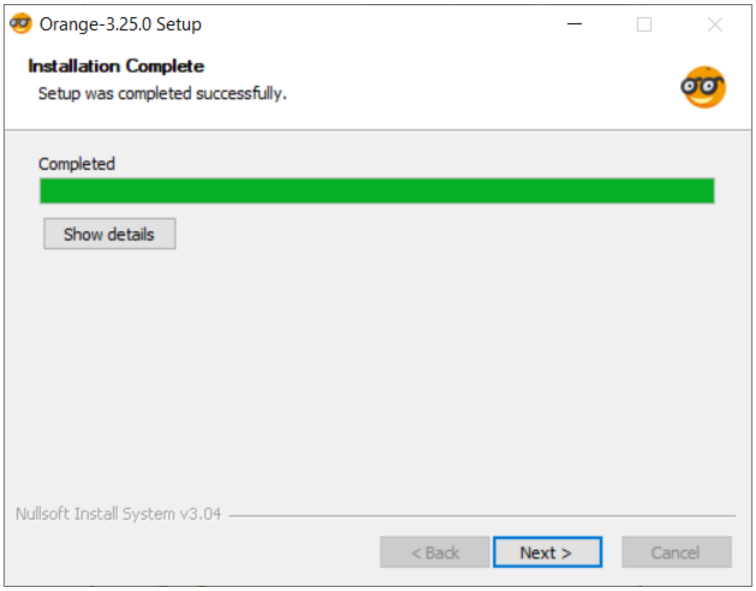

Orange-3.25.0 Setup

Installing

Please wait while Orange-3.25.0 is being installed.

Installing packages (this might take a while)

Progress Bar – let the installation run to completion

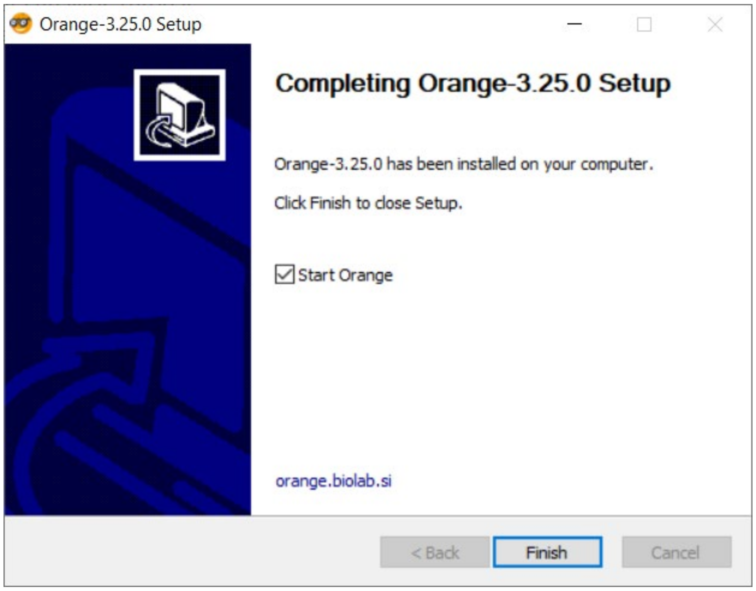

Orange-3.25.0 Setup

Installation Complete

Setup was completed successfully.

Completed

Installation Complete – click **Next**



Completing Orange-3.25.0 Setup

Orange-3.25.0 has been installed on your computer.

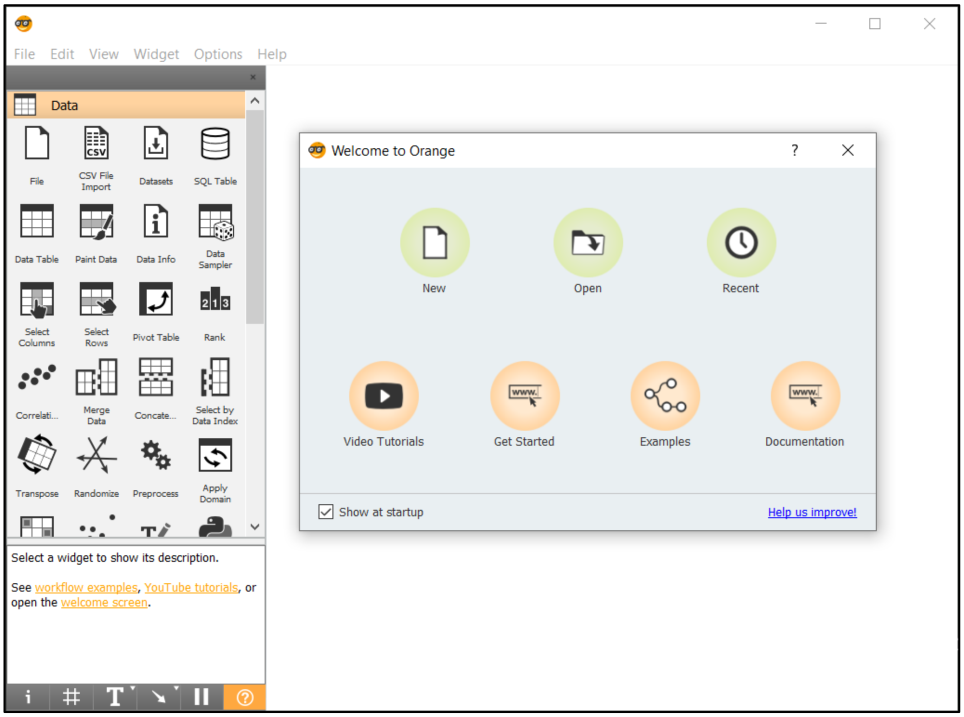
Click Finish to close Setup.

☑ Start Orange

orange.biolab.si

Completion Screen – select **Finish**

Once you select Finish, Orange will launch.



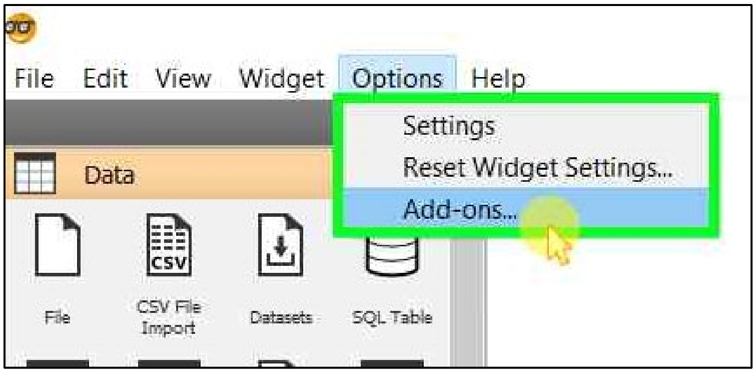
☑ Show at startup

Help us improve!

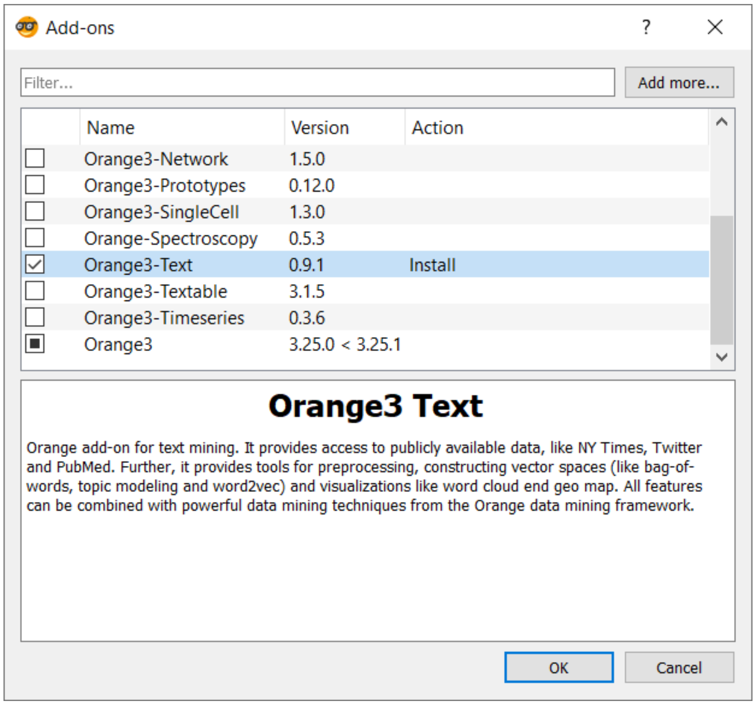
Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

Orange first time launch – **Uncheck show at startup and X out of the Welcome window**. You may also see small prompts about viewing video tutorials and opting in to anonymous data collection. Answer those prompts in whichever manner you wish.



Add-ons – **Click on the Options menu and select Add-ons**…



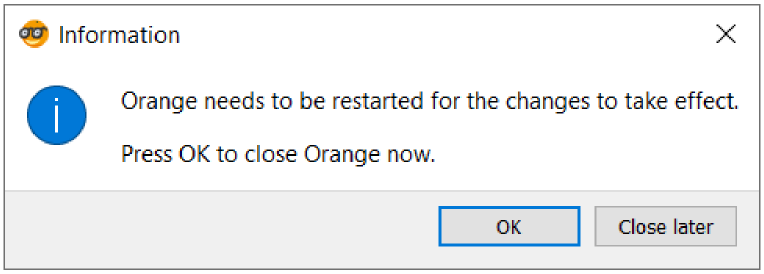
	Name	Version	Action
□	Orange3-Network	1.5.0	
□	Orange3-Prototypes	0.12.0	
□	Orange3-SingleCell	1.3.0	
□	Orange-Spectroscopy	0.5.3	
☑	Orange3-Text	0.9.1	Install
□	Orange3-Textable	3.1.5	
□	Orange3-Timeseries	0.3.6	
▣	Orange3	3.25.0 < 3.25.1	

Orange3 Text

Orange add-on for text mining. It provides access to publicly available data, like NY Times, Twitter and PubMed. Further, it provides tools for preprocessing, constructing vector spaces (like bag-of-words, topic modeling and word2vec) and visualizations like word cloud end geo map. All features can be combined with powerful data mining techniques from the Orange data mining framework.

Text Add-on – Scroll down to the **Orange3-Text add-on, check the box, and click on OK**

You will see a progress bar as the add-on is installed (it may take a while). When the installation is completed you will see a dialog asking you to restart Orange.


Information

Orange needs to be restarted for the changes to take effect.

Press OK to close Orange now.

Orange restart

This is **not** an automatic restart. You must close Orange and reopen the program.

This completes the "Installation Orange and Text Add-on" section of the tutorial.

Pre-Processing Your Data before Using Orange

Data preparation will be 80-90% of your workflow leading up to using Orange for analysis. The old adage of "garbage in equals garbage out" pertains to any project using data and tweets are no exception.

For you to be able to interpret the results of any type of textual analysis of your data, you need to become intimately familiar with the content, regardless of the size of the dataset. This does not mean that you necessarily have to read every tweet in a 10,000+ tweet dataset, but you should sample the data for a set to read closely and choose ways to categorize your data for more comparative analyses. This categorization of your data will help you to answer your research questions.

An example of the decision making and cleanup process of dataset preparation may help you to think about how you can apply these concepts to your own dataset.

Description of Sample Dataset

The example that will be used throughout this tutorial is a dataset collected for a project where we[1] were interested in examining the concept of "contact zones" in a digital interface such as Twitter.[2] We wanted to focus on groups with opposing political ideologies and used #COVID19 and @realDonaldTrump as our search criteria. We selected the hashtag because it was a commonly used, neutral term for the virus (as opposed to misleading and racially charged hashtags such as #chinavirus). It is also more specific to use the scientific label COVID-19 over "coronavirus,"

[1] Emily Johnson, PhD, Assistant Professor of Games and Interactive Media, and Amy Larner Giroux, PhD, Associate Director, Center for Humanities & Digital Research, collected and processed the dataset used in this tutorial.

[2] Mary Louise Pratt, used the term "contact zone" to describe the interactions of groups across different cultures and power disparity. Mary Louise Pratt. "Arts of the Contact Zone." *Profession* (1991), 33–40.

because there are many known coronaviruses. We selected tweets that also tagged the president of the United States during this pandemic because we predicted there would be a dichotomy of opinion that utilized these two markers: those who support the president's actions during this pandemic and those who oppose. The original dataset contained 300,935 tweets between 11 February 2020 and 31 March 2020. To focus on our research questions, we reduced our dataset to those tweets which were retweeted 100 times or more. This reduced our dataset to 1,121 tweets.

To classify and categorize the dataset in a meaningful way for our research, we looked at the specific users who sent these tweets and the dataset was comprised of 543 unique individuals. Twitter profile data for these individuals was collected and matched with the tweets in the dataset. The general content of each user's tweets appearing in the dataset and number of followers were used to categorize the users in three ways: as influencers, as opinion-leaders, and by their political leaning (left, right, or neutral). A user's overall categorization was then assigned to their individual tweets.

Influencers were determined to be users who have 10,000 followers or more. Users were scored as opinion-leaders if most of their tweets included information or a link to an article or video. For example, Donald Trump tweeted on March 19, "America's Private Sector is stepping up to help us be STRONG! Many of the Nation's distillers, large and small, are producing and donating hand sanitizer to help fight #COVID19. THANK YOU! https://www.nbcnews.com/news/us-news/distilleries-using-high-proof-alcohol-make-hand-sanitizer-n1161371." This tweet includes information, "distillers, large and small, are producing and donating hand sanitizer" as well as a link to a news report. The veracity of the information was not scored—as long as it was written such that an uninformed follower could interpret it as information, it was counted as such. Therefore, any given participant could be scored as both an influential and an opinion-leader.

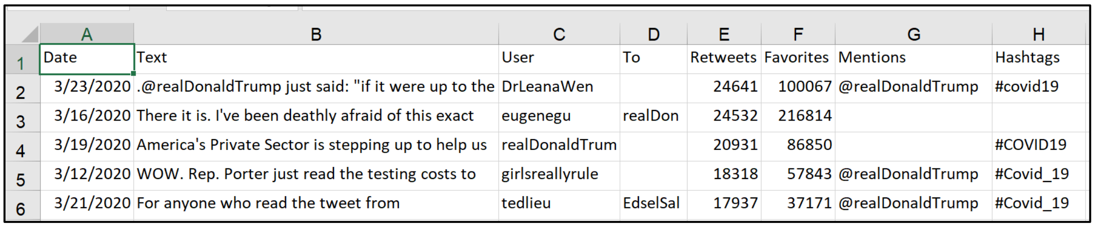
Political leaning was determined by examining the user profile and general contents of that user's tweets within the dataset. Any profile whose tweets were not clearly supportive or critical of either side of the political divide were scored as "neutral."

Excel Tips for Dataset Preparation

For basic file preparation, please see the "Open the Dataset in Excel and Prepare it for Analysis" section of the Twitter Data Scraping Tutorial. Following the directions in that tutorial will help you create a base level Excel spreadsheet with your tweet data. The following list of tips will use the COVID19 dataset as its base.

You are not expected to complete this level of data preparation on your own data for this tutorial. Read through the information to understand it in terms of the sample dataset but do not try to perform these steps with your own data until after the Institute is over.

Some of the columns that are created during the scraping process (e.g. count, ID) are not needed for analysis. The columns that should be retained for use in Orange are Date, Text, User, To, Retweets, Favorites, Mentions, and Hashtags as seen below.



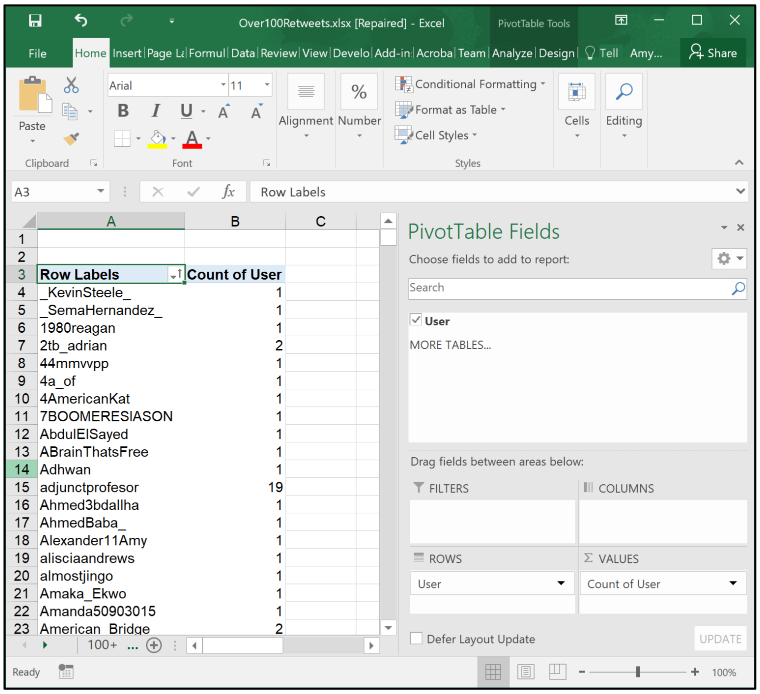
	A	B	C	D	E	F	G	H
1	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags
2	3/23/2020	.@realDonaldTrump just said: "if it were up to the	DrLeanaWen		24641	100067	@realDonaldTrump	#covid19
3	3/16/2020	There it is. I've been deathly afraid of this exact	eugenegu	realDon	24532	216814		
4	3/19/2020	America's Private Sector is stepping up to help us	realDonaldTrum		20931	86850		#COVID19
5	3/12/2020	WOW. Rep. Porter just read the testing costs to	girlsreallyrule		18318	57843	@realDonaldTrump	#Covid_19
6	3/21/2020	For anyone who read the tweet from	tedlieu	EdselSal	17937	37171	@realDonaldTrump	#Covid_19

Columns for use in Orange

You can create multiple tabs (worksheets) within your Excel file to hold the raw data (the data from the scraper) and a copy of that data for categorization.

Pivot Tables

Pivot tables are a handy way to filter and examine data. The image below shows a pivot table for the User column of tweet data and the number of tweets from the dataset (Count of User – since the user name appears on each row of tweets).



	A	B
3	Row Labels	Count of User
4	_KevinSteele_	1
5	_SemaHernandez_	1
6	1980reagan	1
7	2tb_adrian	2
8	44mmvvpp	1
9	4a_of	1
10	4AmericanKat	1
11	7BOOMERESIASON	1
12	AbdulElSayed	1
13	ABrainThatsFree	1
14	Adhwan	1
15	adjunctprofesor	19
16	Ahmed3bdallha	1
17	AhmedBaba_	1
18	Alexander11Amy	1
19	alisciaandrews	1
20	almostjingo	1
21	Amaka_Ekwo	1
22	Amanda50903015	1
23	American_Bridge	2

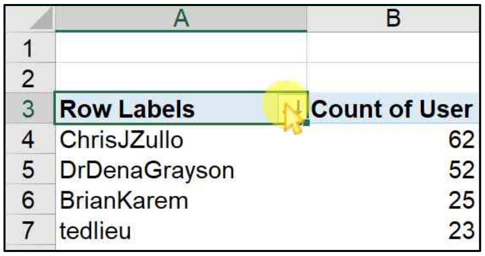
☑ User

ROWS: User

VALUES: Count of User

Pivot table of unique users sorted alphabetically (default)

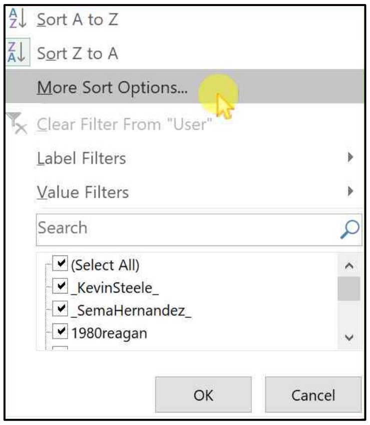
The default sort order for the pivot table is alphabetical. We can change the sort to be by the number of tweets (count of user) by **clicking on the arrow next** to the Row Labels in the image below.



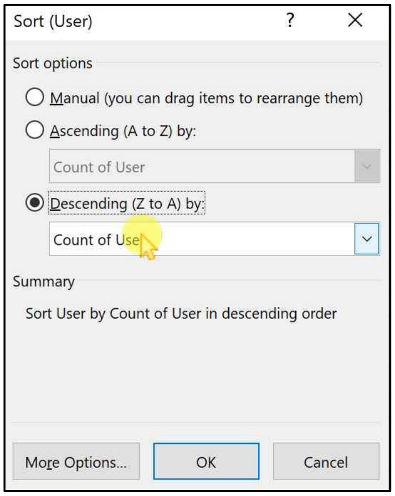
	A	B
1		
2		
3	Row Labels	Count of User
4	ChrisJZullo	62
5	DrDenaGrayson	52
6	BrianKarem	25
7	tedlieu	23

Using Custom Sort (down arrow next to Row Labels) to sort by number of tweets

After clicking on the **down arrow**, choose **More Sort Options…**



Sort Selection – Choose **More Sort Options**



Sort (User)

Sort options

○ Manual (you can drag items to rearrange them)

○ Ascending (A to Z) by:

Count of User

● Descending (Z to A) by:

Count of User

Summary

Sort User by Count of User in descending order

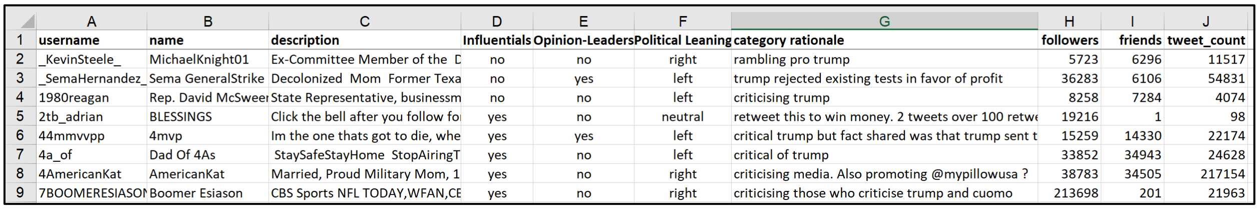
Sort Options – Choose **Descending** by **Count of User**

By sorting the pivot table in this manner, you can examine which users had the most tweets within the dataset.

VLOOKUP

Depending upon your decisions for categorizing your data, you may reach a point where you have categorical data on one tab and you main tweet information on another. Using the VLOOKUP function in Excel is a good way to merge the two datasets. It allows you to match columns between tabs and copy data from one tab to another.

Continuing with our sample dataset, the image below shows part of the information from our categorization process. To use the VLOOKUP function, the first column **must** contain the data field on which you will match (in this case username).



	A	B	C	D	E	F	G	H	I	J
1	**username**	**name**	**description**	**Influentials**	**Opinion-Leaders**	**Political Leaning**	**category rationale**	**followers**	**friends**	**tweet_count**
2	_KevinSteele_	MichaelKnight01	Ex-Committee Member of the D	no	no	right	rambling pro trump	5723	6296	11517
3	_SemaHernandez_	Sema GeneralStrike	Decolonized Mom Former Texa	no	yes	left	trump rejected existing tests in favor of profit	36283	6106	54831
4	1980reagan	Rep. David McSweer	State Representative, businessm	no	no	left	criticising trump	8258	7284	4074
5	2tb_adrian	BLESSINGS	Click the bell after you follow fo	yes	no	neutral	retweet this to win money. 2 tweets over 100 retwe	19216	1	98
6	44mmvvpp	4mvp	Im the one thats got to die, whe	yes	yes	left	critical trump but fact shared was that trump sent t	15259	14330	22174
7	4a_of	Dad Of 4As	StaySafeStayHome StopAiringT	yes	no	left	critical of trump	33852	34943	24628
8	4AmericanKat	AmericanKat	Married, Proud Military Mom, 1	yes	no	right	criticising media. Also promoting @mypillowusa ?	38783	34505	217154
9	7BOOMERESIASOI	Boomer Esiason	CBS Sports NFL TODAY,WFAN,CE	yes	no	right	criticising those who criticise trump and cuomo	213698	201	21963

Categorization of users as influential, opinion-leaders, and political leaning

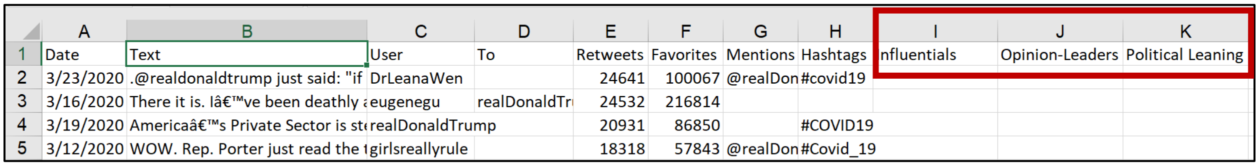
The other columns in the above image contain data from the user profile (columns B, C), the categorization we assigned to that specific user (columns, D, E, F), our rationale for the categories (column G), and the counts from the user profile (columns H, I, J).

Naming the tabs something other than the default "Sheet 1" will make the VLOOKUP formula more readable. Double-clicking on the tab name will allow you to change it. In the dataset used in this example, the profile information in the above image is on a tab named CodedProfiles.

We want to copy the category columns (D, E, F) back to our main data spreadsheet using VLOOKUP, which has four parameters in the formula:

1. The cell containing the term to look up. In this case we are using column C from our tweets tab, which contains the username.
2. The range of data in which to search. In our case the CodedProfiles tab
3. The number of the column to return. This is one-based (A = 1, B = 2, etc.) and for our columns D, E, and F, these would be 4, 5, and 6.
4. Whether to use a close match (TRUE) or an exact match (FALSE). We will use exact (FALSE).

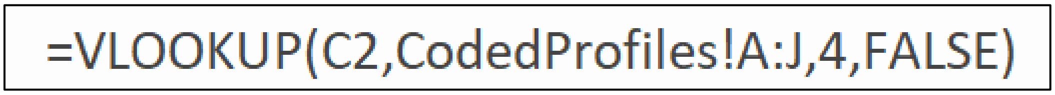
Add three empty columns to the main tweet tab to hold the categorization data.



	A	B	C	D	E	F	G	H	I	J	K
1	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags	nfluentials	Opinion-Leaders	Political Leaning
2	3/23/2020	.@realdonaldtrump just said: "if	DrLeanaWen		24641	100067	@realDon	#covid19			
3	3/16/2020	There it is. lâ€™ve been deathly	eugenegu	realDonaldTr	24532	216814					
4	3/19/2020	Americaâ€™s Private Sector is st	realDonaldTrump		20931	86850		#COVID19			
5	3/12/2020	WOW. Rep. Porter just read the	girlsreallyrule		18318	57843	@realDon	#Covid_19			

Three new, empty columns for categories

In the first cell of the Influentials column we enter the following formula:

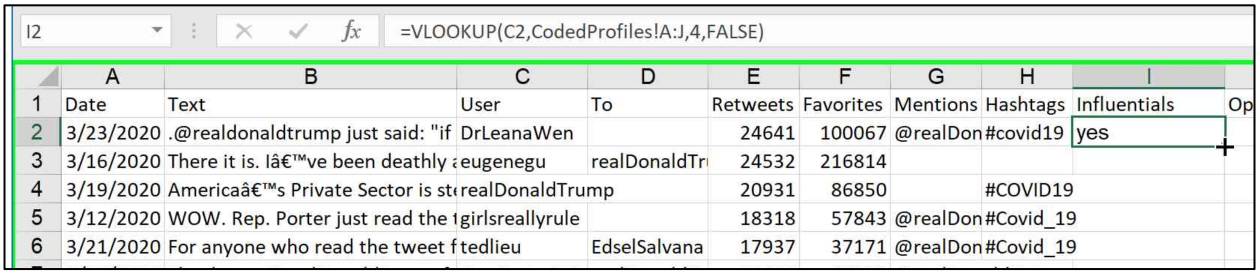


=VLOOKUP(C2,CodedProfiles!A:J,4,FALSE)

VLOOKUP example

C2 equates to DrLeanaWen, the username on that row; CodedProfiles!A:J is the categorization tab contents (all columns); 4 is the fourth column (Influentials on the CodedProfiles tab), and FALSE is for an exact match.

The formula set Influentials to "yes" for DrLeanaWen's tweet as shown below. To replicate this formula down the column, hover your cursor over the right-lower corner of the cell to get the + cursor and then double-click.



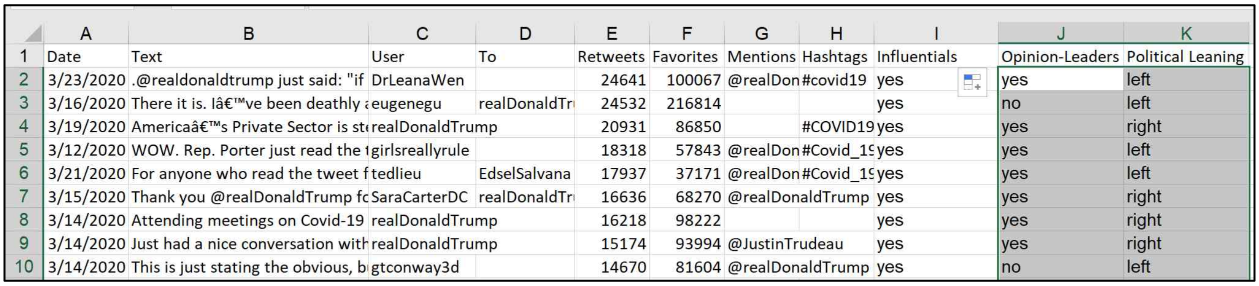
I2

=VLOOKUP(C2,CodedProfiles!A:J,4,FALSE)

	A	B	C	D	E	F	G	H	I	
1	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags	Influentials	Op
2	3/23/2020	.@realdonaldtrump just said: "if	DrLeanaWen		24641	100067	@realDon	#covid19	yes	
3	3/16/2020	There it is. Iâ€™ve been deathly a	eugenegu	realDonaldTr	24532	216814				
4	3/19/2020	Americaâ€™s Private Sector is ste	realDonaldTrump		20931	86850		#COVID19		
5	3/12/2020	WOW. Rep. Porter just read the t	girlsreallyrule		18318	57843	@realDon	#Covid_19		
6	3/21/2020	For anyone who read the tweet f	tedlieu	EdselSalvana	17937	37171	@realDon	#Covid_19		

Entered formula with cursor in position to replicate down the column

After copy the VLOOKUP formula to the Opinion-Leaders and Political Leaning columns, we used the same procedure to replicate the formulas down the columns.



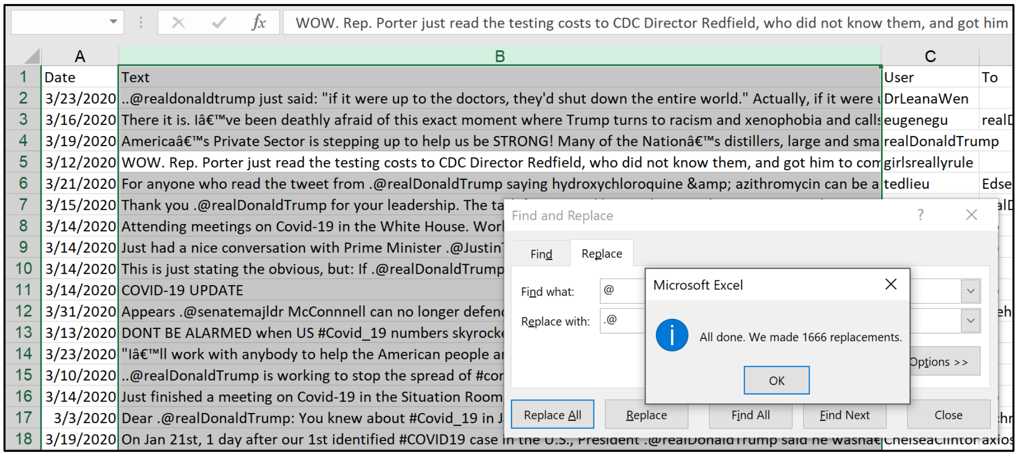
	A	B	C	D	E	F	G	H	I	J	K
1	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags	Influentials	Opinion-Leaders	Political Leaning
2	3/23/2020	.@realdonaldtrump just said: "if	DrLeanaWen		24641	100067	@realDon	#covid19	yes	yes	left
3	3/16/2020	There it is. Iâ€™ve been deathly a	eugenegu	realDonaldTr	24532	216814			yes	no	left
4	3/19/2020	Americaâ€™s Private Sector is ste	realDonaldTrump		20931	86850		#COVID19	yes	yes	right
5	3/12/2020	WOW. Rep. Porter just read the t	girlsreallyrule		18318	57843	@realDon	#Covid_19	yes	yes	left
6	3/21/2020	For anyone who read the tweet f	tedlieu	EdselSalvana	17937	37171	@realDon	#Covid_19	yes	yes	left
7	3/15/2020	Thank you @realDonaldTrump fo	SaraCarterDC	realDonaldTr	16636	68270	@realDonaldTrump		yes	yes	right
8	3/14/2020	Attending meetings on Covid-19	realDonaldTrump		16218	98222			yes	yes	right
9	3/14/2020	Just had a nice conversation with	realDonaldTrump		15174	93994	@JustinTrudeau		yes	yes	right
10	3/14/2020	This is just stating the obvious, b	gtconway3d		14670	81604	@realDonaldTrump		yes	no	left

Replicated formulas

Find and Replace

The "@" sign at the beginning of a cell in Excel causes the program to think that it is a formula instead of text. When doing a find and replace on a column such as the tweet Text (B), Excel will stop processing when it encounters an initial @ and give you an error.

There are a couple of choices to alleviate this issue. You can replace the "@" with an empty string and therefore remove it from all places, but if you want to keep it in the text for display, you will need to replace it with something else. By using something such as .@ (a period before the @) so that it stops it from thinking it is a formula, it still gives the visual of having the @.



WOW. Rep. Porter just read the testing costs to CDC Director Redfield, who did not know them, and got him

	A	B	C	
1	Date	Text	User	To
2	3/23/2020	..@realdonaldtrump just said: "if it were up to the doctors, they'd shut down the entire world." Actually, if it were u	DrLeanaWen	
3	3/16/2020	There it is. Iâ€™ve been deathly afraid of this exact moment where Trump turns to racism and xenophobia and calls	eugenegu	realD
4	3/19/2020	Americaâ€™s Private Sector is stepping up to help us be STRONG! Many of the Nationâ€™s distillers, large and sma	realDonaldTrump	
5	3/12/2020	WOW. Rep. Porter just read the testing costs to CDC Director Redfield, who did not know them, and got him to com	girlsreallyrule	
6	3/21/2020	For anyone who read the tweet from .@realDonaldTrump saying hydroxychloroquine &amp; azithromycin can be a	tedlieu	Edse
7	3/15/2020	Thank you .@realDonaldTrump for your leadership. The ta		alD
8	3/14/2020	Attending meetings on Covid-19 in the White House. Worl		
9	3/14/2020	Just had a nice conversation with Prime Minister .@JustinT		
10	3/14/2020	This is just stating the obvious, but: If .@realDonaldTrump		
11	3/14/2020	COVID-19 UPDATE		
12	3/31/2020	Appears .@senatemajldr McConnnell can no longer defend		eh
13	3/13/2020	DONT BE ALARMED when US #Covid_19 numbers skyrocke		
14	3/23/2020	"Iâ€™ll work with anybody to help the American people ar		
15	3/10/2020	..@realDonaldTrump is working to stop the spread of #cor		
16	3/14/2020	Just finished a meeting on Covid-19 in the Situation Room		
17	3/3/2020	Dear .@realDonaldTrump: You knew about #Covid_19 in J		hr
18	3/19/2020	On Jan 21st, 1 day after our 1st identified #COVID19 case in the U.S., President .@realDonaldTrump said he wasnâ€	ChelseaClinton	axios

Find and Replace

Find what: @

Replace with: .@

Microsoft Excel

All done. We made 1666 replacements.

Replacement of @ – **Find** and **Replace @** with **.@**

Sometimes there are encoding problems in the tweets where characters such as an apostrophe come into your dataset as â€™ as you can see in lines 3 and 4 in the above image. The reason for the replacement of @ is to make the data fixes for encoding problems easier as you will not get formula errors. The table below shows some of the encoding issues found in the sample dataset. This is not all inclusive as other datasets may include different encoding issues.

Character	**Bad Encoding**
Apostrophe	â€™
Left double quote	â€œ
Right double quote	â€
HTML equivalent for ampersand	&amp;
HTML for >	&gt;
HTML for <	&lt;

Using Ctrl-H (Find/Replace) you can reset these to the correct character.

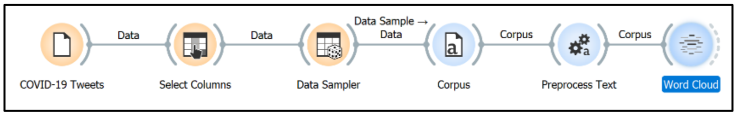
At this point, the data has been categorized and cleaned of unwanted encoding issues. It now is ready to be imported into Orange.

Remember that the above steps are not intended to be done to your data prior to the workshops at the NEH Digital Culture Institute. It was important to document the steps for you to follow at a later date.

This completes the "Pre-Processing Your Data before Using Orange" section of the tutorial.

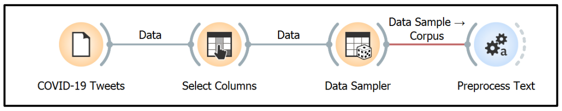
Overall Orange Tips

Each widget within the workflow is connected to the next in the process. The lines displayed between the widgets tells you the type of input/output for the step in the process. For example, "data" is the type used from the File through the Corpus widget in the example below. Once the process passes the Corpus widget, the type of data is a Corpus.



Orange workflow example

If you try to connect widgets that do not have the same input type requirements, you will see the connecting line turn red as in the example below where the output of the Data Sampler (data) cannot be used for input to Preprocess Text as it expects the type of Corpus.

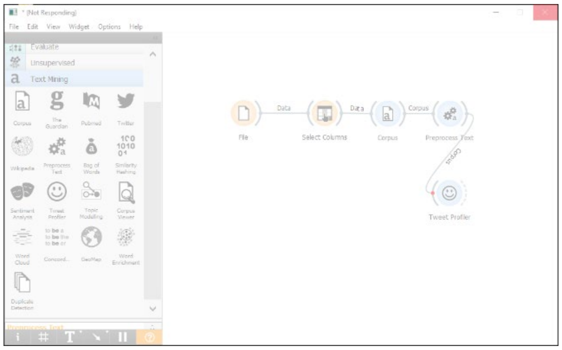


Workflow type mismatch error

Hovering your mouse over a widget will show the input/output criteria. Additionally, clicking once on a widget to highlight it and then pressing F1 will bring up the detailed help information for the widget.

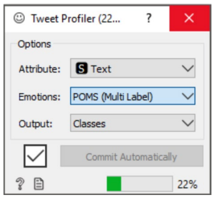
Be Patient

Depending on your dataset size, the processing time for various steps in an Orange workflow can take time. The default for all of the widgets is to process (commit) automatically when they are connected into a workflow. Sometimes the program, as shown below, will show "Not responding" in the title bar as it is processing. Note the red dot at the Tweet Profiler widget which shows that it is processing the data in that widget. Be patient and let the processes finish if you get a not responding message.



Not responding example

You can control when a widget runs, for example the Tweet Profiler shown below, by unchecking the "Commit Automatically" option within the specific widget. If you uncheck this box, you will need to click on the button to make the widget process.



Tweet Profiler (22... ?

Options

Attribute: S Text

Emotions: POMS (Multi Label)

Output: Classes

☑ Commit Automatically

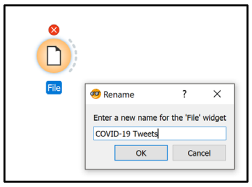
22%

Tweet Profiler dialog with "Commit Automatically" and progress bar

Most of the processes will also show a percentage complete bar at the bottom of the dialog box. Sometimes, like the "not responding" message, this bar does not update consistently until the process is complete.

Renaming Widgets

Each widget can be renamed. Right-click and pick Rename, or hit F2 when the widget it active. Being able to rename widgets helps to keep your workflow manageable.



Rename ? ×

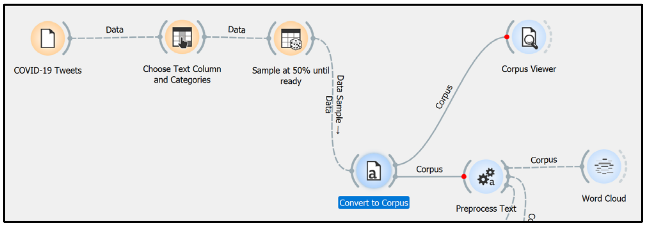
Enter a new name for the 'File' widget

COVID-19 Tweets

OK Cancel

Renaming the File widget

You can use the widget names to note what you are doing with each. In the image below, the Select Columns widget was renamed to "Choose Text Column and Categories" and the Data Sampler widget was renamed to "Sample at 50% until ready." Notes like these can help remind you of things to go back and do, such as putting the sampling back to 100% once your full process is ready.



Renamed widgets for processing notes

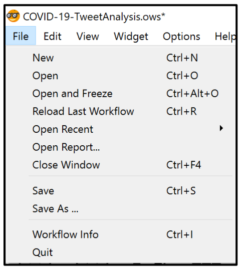
Opening Workflow and Data Files

Saved workflows

There are three ways to open a saved workflow.

1. Double-click on the .ows file in File Explorer
2. Open Orange and use Ctrl-O or File -> Open (see image below)
3. Open Orange and use Ctrl-Alt-O or File -> Open and Freeze (see image below)

Opening a saved workflow with either of the first two options will immediately begin the data flow through the widgets. As noted before in this tutorial, you may get a "not responding" message until all of the processes are completed. The third option, Open and Freeze will keep the internal flow processes from starting immediately.



Orange File menu

Opening data files

When you use the File widget within a workflow, the widget looks for specific file types allowed as input. It has no understanding of the Windows Desktop, or other folders in the explorer window for opening files. If your data is on your Desktop, you will need to navigate to **Windows -> Users -> your user name -> Desktop** to finally be able to open files or folders in that place.

Widget Notes

A couple of other caveats before moving into the examples of creating workflows.

1. There are not a lot of Finish or OK buttons on various screens within Orange. Sometimes you will leave a window open while you do other tasks and watch how those tasks affect your output. There will be comments in the tutorial text that will let you know when it is safe to close a window.
2. There are multiple ways to add widgets to a workflow, such as double or right-clicking on the work area or selecting from the widget window. This tutorial will use the widget window but you are welcome to use the other methods.

This completes the "Overall Orange Tips" section of the tutorial

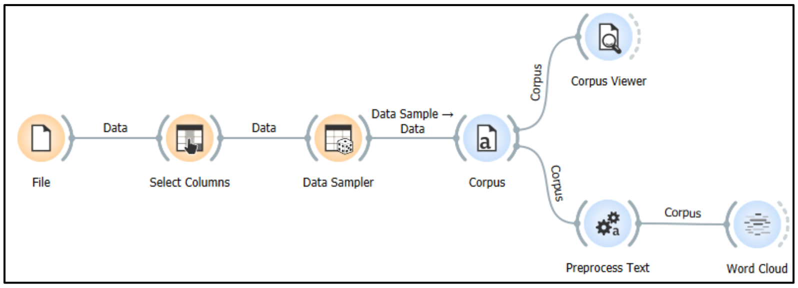
First Workflow -> File Load to Word Cloud

From the Twitter Data Scraping tutorial, or some other method of data collection, you should have an Excel spreadsheet containing tweet data.

If you are following this tutorial as an attendee of the NEH Digital Culture Institute, for this first workflow, please use the tweet data that was supplied to you as it includes data that

interests you. Open the CSV file in Excel, delete the first column (the count), the second column (ID) and save the file as an Excel spreadsheet (.xslx).

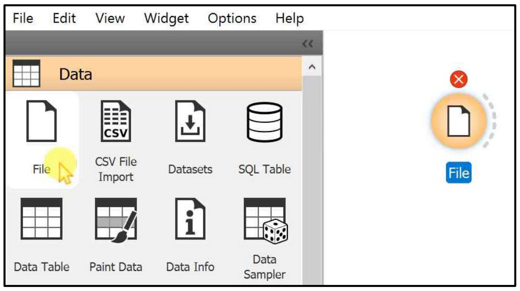
The image below is the workflow we will build during this part of the tutorial. Each widget will be described in order. Open Orange and start with an empty work area.



First workflow

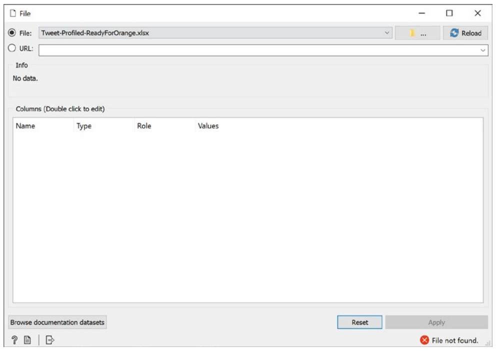
File Load

Click on **File** in the Data section of the widget window to add a File widget to the work area.



Adding a File widget

Double-click on the File widget to open it. If you have opened files in other workflows, the last file may be shown. In the bottom right of this window it will show whether that file still exists. In this example the Tweet-Profiled-ReadyForOrange spreadsheet was not found.



● File: Tweet-Profiled-ReadyForOrange.xlsx

○ URL:

Info

No data.

Columns (Double click to edit)

Name Type Role Values

File not found.

File selection dialog