Workbook 2018

# Introduction to KNIME Analytics

## ANALYSING DATA USING KNIME

**Files used in this workshop (along with download links):**

- MyFirstTestWorkflow.zip (https://bit.ly/2IKOzuI)
  - 2018-03Mar.csv: (https://bit.ly/2jXiMZl)
- MyFirstWorkflowForCrimeAPI.zip: (https://bit.ly/2rJlZzT)
- MyFirstTwitterWorkflow.zip: (https://bit.ly/2L4hGY6)
- MySecondTwitterWorkflow.zip: (https://bit.ly/2IptuCj)
  - sentiments.csv: (https://bit.ly/2wJF6yH)
  - dictionary-partsofspeech.txt: (https://bit.ly/2IDEC1U)

**YOU CAN ALSO DOWNLOAD THIS WORKBOOK**
**https://bit.ly/2rH7NYe**

# 1 Introduction to Konstanz Information Miner (KNIME)

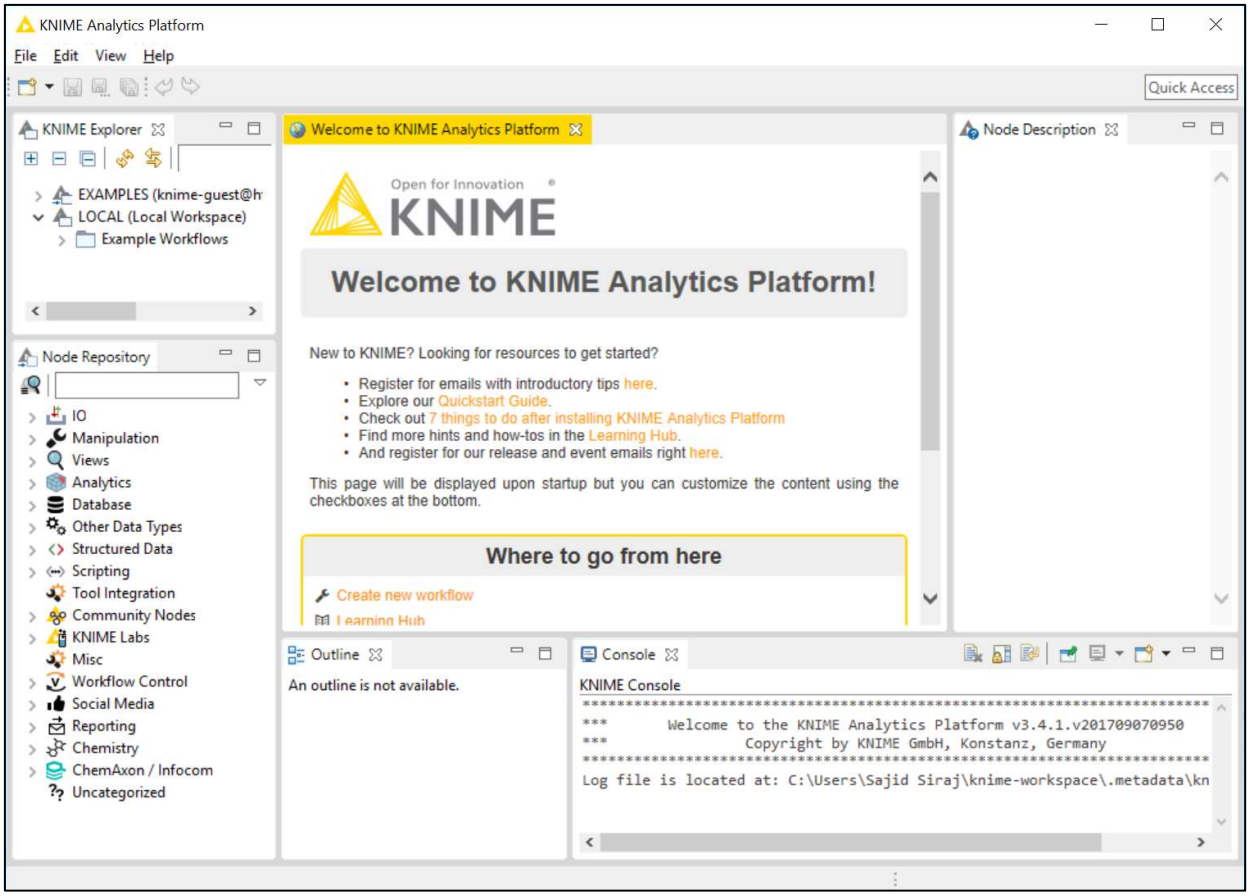
Our aim is to become familiar with workflow-based analytics and to performing some common data analytics tasks. We will be using KNIME Analytics software for this purpose. Once you are comfortable in using KNIME, you can easily migrate to any other workflow-based tool, for example, RapidMiner, SAS Analytics etc.

**N.B. We are using KNIME version 3.3, if you downloading the latest version (3.5+), please ensure that you download all extensions (File Menu -> Install KNIME Extensions).**

## 1.1 ACTIVITY 1 OPENING KNIME ANALYTICS PLATFORM

►**Open KNIME from the start menu**, or simply double-click the knime.exe file from the folder where you extracted the full version of KNIME (it is open-source, you can download it for free).

On opening it first time, you might be asked to choose the folder where you will save your workflows. This folder is called "Local Workspace" by KNIME. All the folders and files saved inside your workspace will appear on the top-left window of KNIME user interface (see below).



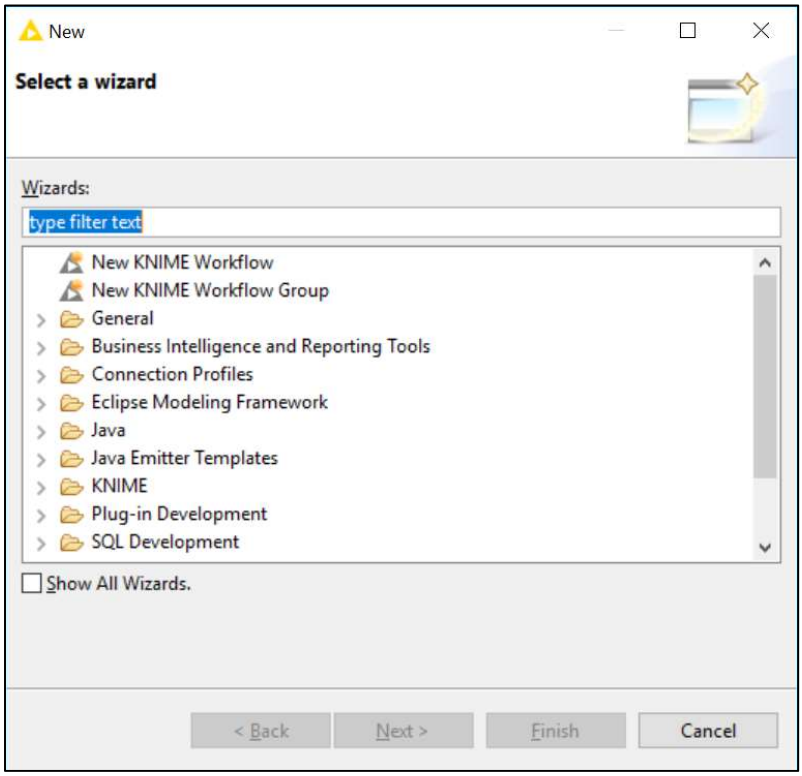

The bottom-left window is the most important one to develop your analytics workflow. This window contains the "Node Repository". Each node is a process that you will apply to your data, it can be as simple as doing arithmetic to as complicated as deep learning neural networks or training support vector machines.

## 1.2 ACTIVITY 2 CREATING WORKFLOW

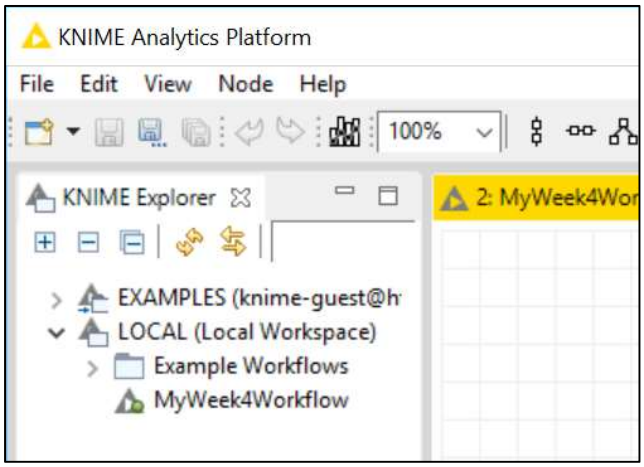
Create a new workflow by choosing "New…" from the File menu on the top-left of screen.

**►Select "New KNIME Workflow" and click next.**
It will ask you to provide a name for this new workflow, you can name anything you prefer but it is better to keep it meaningful.

**►Type the name as "MyFirstWorkflow"**
Feel free to choose anything that you find appropriate.



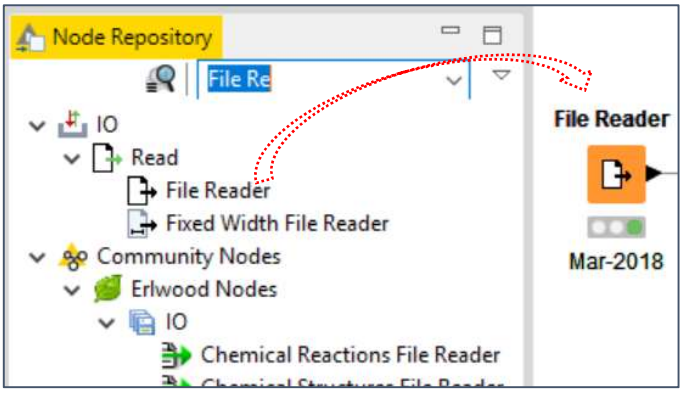

After creating the workflow, you will notice the blank canvas in the central window and also the name of this new workflow appearing on the workspace hierarchy at top-left corner, as shown below:




We are now ready to drag and drop any nodes from the bottom-left window.

## 1.3 ACTIVITY 3 ACCESSING COMMA-SEPARATED DATA (CSV) FILE

►Find the node "File Reader" inside the **Read** folder which is available under the **IO** group of nodes.




►Drag this node to your workflow in the centre.

KNIME will place a square block with title File Reader and will put a label "Node 1" to this instance. You can rename this instance to Mar-2018.

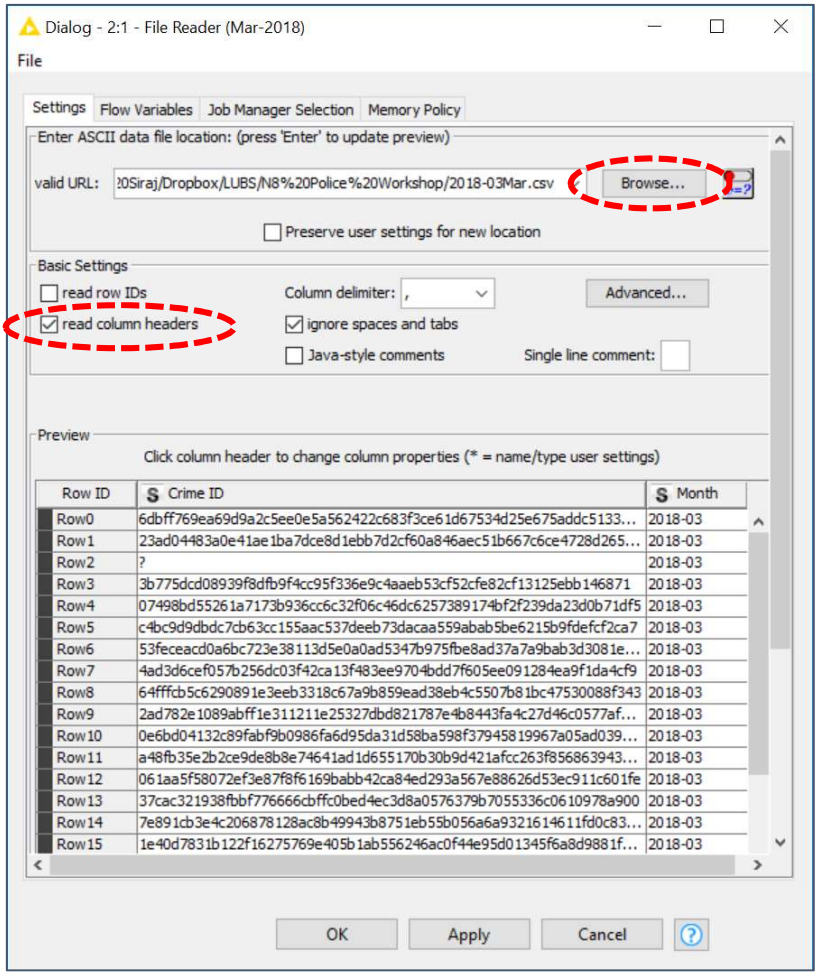
You can place same nodes many times, there is virtually no limit on the number of copies you create for each node.

►Double-click on this newly created node.

You will see a configuration window popping out of this node. This will be used to configure the file path and settings to load data for further analysis.

►Browse and select the file which you can download from the link provided earlier.

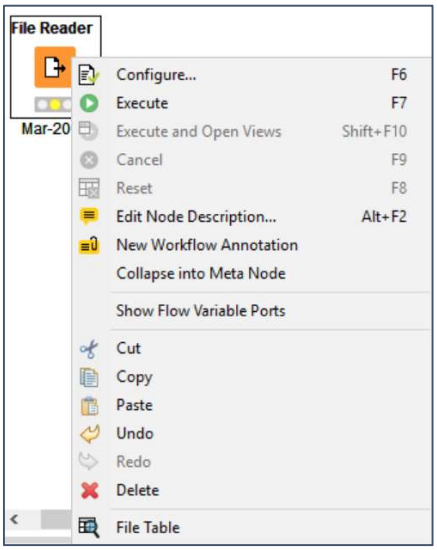
Note that the "read column headers" setting must be checked for this file which implies that the top row in the file is the header row. This is important to always check because your clients sometimes provide files without headers (they may provide columns information separately).




►Press OK to close this box.

►Now right-click on the **File Reader** node and press "Execute" command on the pop-up menu.

You will notice that yellow light will turn into green if everything goes smooth. If you fail to load file, KNIME will change yellow light to red. In case of red, you must check your configuration.




►After successfully executing the node, again right-click on it and select “File Table”. This will show you the data loaded from the file.

KNIME is a great tool for investigating each and every step of your analysis. You can inspect all the intermediate outputs generated by each node. This will become clearer as we add more nodes into our workflow.

File Table - 2:1 - File Reader (Mar-2018)

File

Table "2018-03Mar.csv" - Rows: 27674 | Spec - Columns: 12 | Properties | Flow Variables

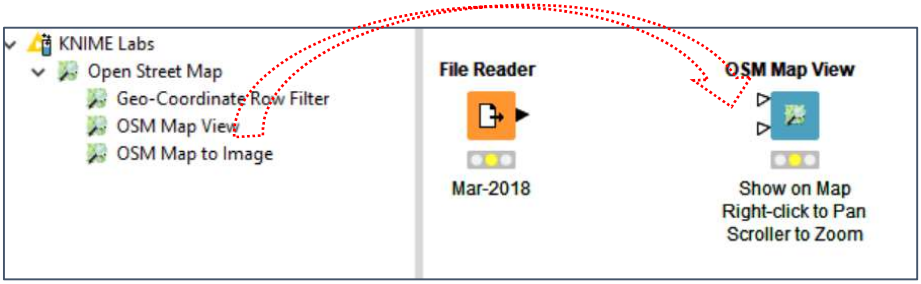
| Row ID | Crime ID | Month | Reported by | Falls within | Longitude | Latitude | Locatio |
|---|---|---|---|---|---|---|---|
| Row0 | 6dbff769ea6... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.765 | 53.535 | On or near |
| Row1 | 23ad04483a0... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.062 | 53.388 | On or near |
| Row2 | ? | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.863 | 53.939 | On or near |
| Row3 | 3b775dcd089... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.874 | 53.943 | On or near |
| Row4 | 07498bd5526... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.879 | 53.944 | On or near |
| Row5 | c4bc9d9dbdc... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.863 | 53.94 | On or near |
| Row6 | 53feceacd0a... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.863 | 53.939 | On or near |
| Row7 | 4ad3d6cef05... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.883 | 53.942 | On or near |
| Row8 | 64fffcb5c629... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.883 | 53.942 | On or near |
| Row9 | 2ad782e1089... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.898 | 53.946 | On or near |
| Row10 | 0e6bd04132c... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.888 | 53.945 | On or near |
| Row11 | a48fb35e2b2... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.822 | 53.925 | On or near |
| Row12 | 061aa5f5807... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.822 | 53.925 | On or near |
| Row13 | 37cac321938... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.822 | 53.925 | On or near |
| Row14 | 7e891cb3e4c... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.826 | 53.92 | On or near |
| Row15 | 1e40d7831b1... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.822 | 53.924 | On or near |
| Row16 | 04cdc791924... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.819 | 53.921 | On or near |
| Row17 | 01b23787bfc... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.79 | 53.916 | On or near |
| Row18 | d868ae82409... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.83 | 53.921 | On or near |
| Row19 | ? | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.797 | 53.92 | On or near |
| Row20 | f7766720a9a... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.801 | 53.92 | On or near |
| Row21 | f4ce2600e73... | 2018-03 | West Yorkshire Police | West Yorkshire ... | -1.812 | 53.924 | On or near |

For today, let us do a simple visual inspection of data.

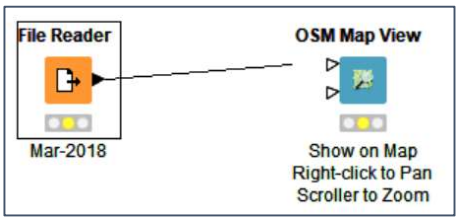
## 1.4 ACTIVITY 4 VISUALISING DATA ON THE MAP

►Look for the search box at the top of the Node Repository window, and type "Street". Alternatively, you can browse the group "KNIME Labs" and find sub-group "Open Street Map".

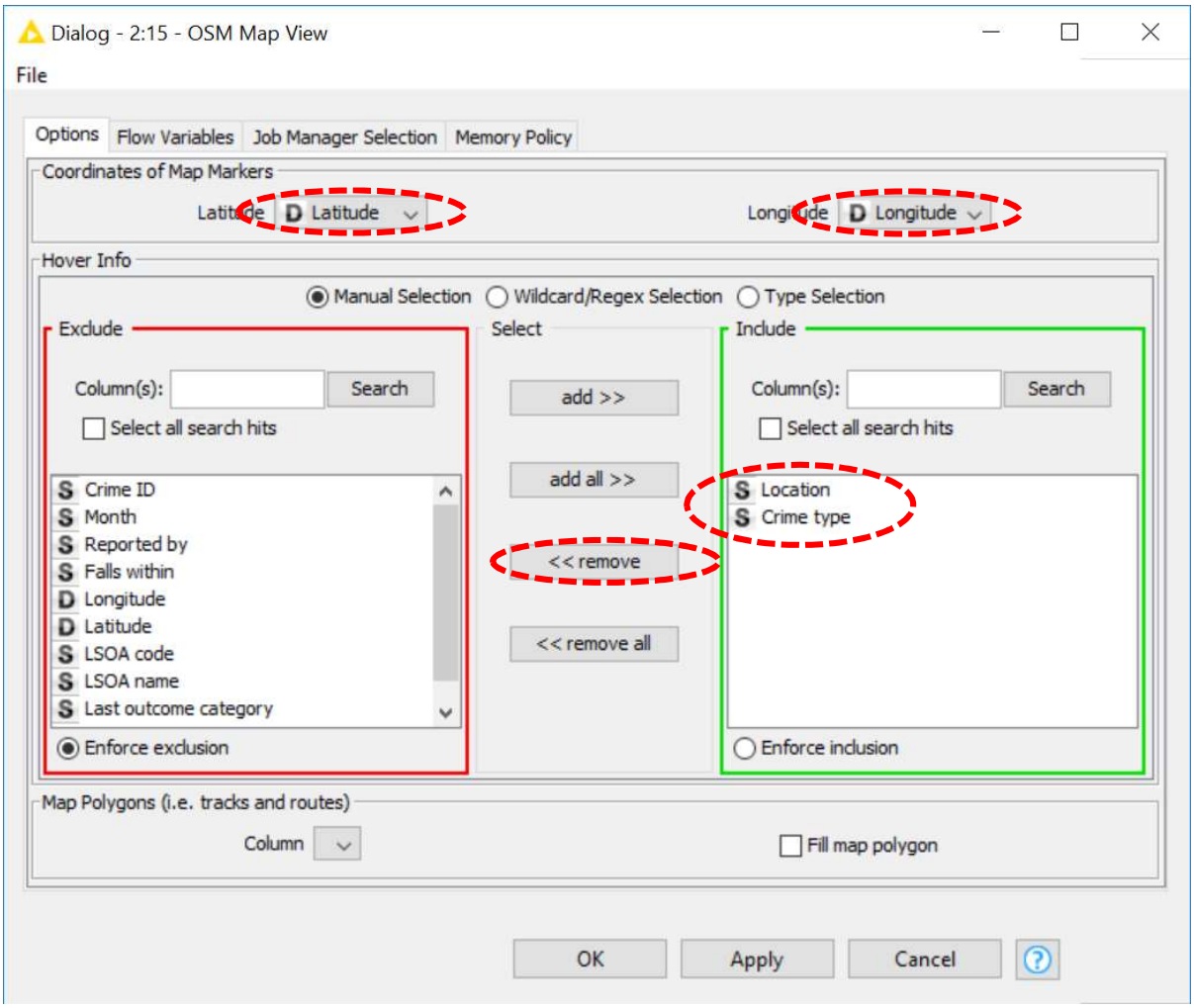
►From the Open Street Map group, select and drag "**OSM Map View**" node on your workflow.




►Click on the right-arrow of "**File Reader**" node and drag it to the left-top arrow of "**OSM Map View**". Ensure that you have connected the link. This implies that output of File Reader will become the first input of OSM Map View.

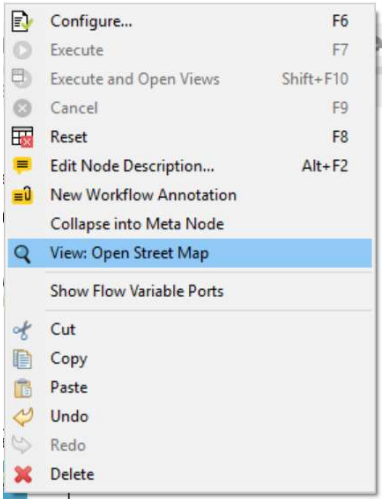


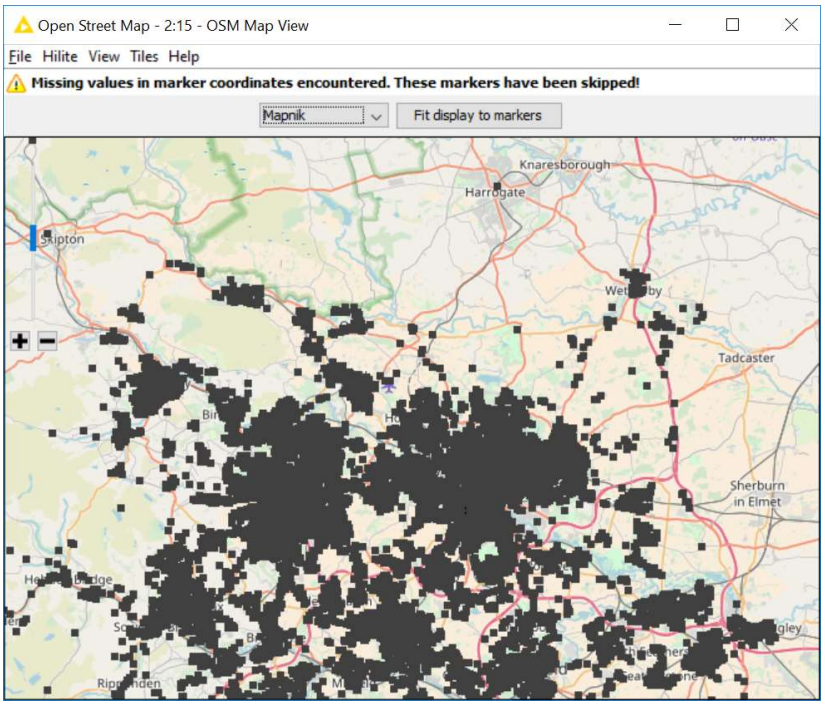

►Now double-click the OSM Map View to configure it.



►As shown in the screenshot on last page, choose the Latitude and Longitude columns carefully. Also, only select Location and Crime Type for the information visible on the map. We are doing this for better readability.

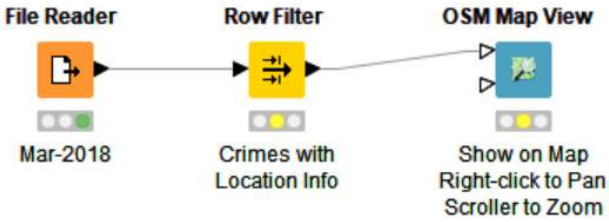
►Click "View: Open Street Map" on the menu for this node. You will notice, the crime cases are shown as block dots in West Yorkshire area. If you cannot see Yorkshire area, please use navigate by right-clicking and dragging the map. Open Street Map is the map of whole world so you need to locate UK and then West Yorkshire yourself.

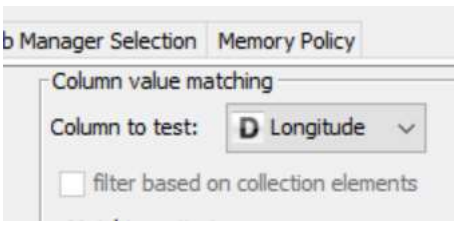


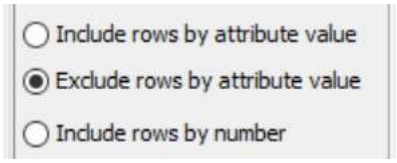



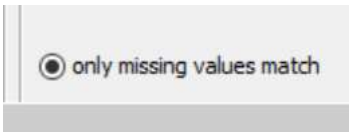

►You will notice the warning that some of the data come without location information. You may wish to remove this warning by adding a filter for selecting only useful data.

►Search for "Row Filter" node and place this node in the middle of File Reader and OSM Map View. Reconnect the wires so that the output of File Reader gets filtered before entering the OSM Map View. Now configure the Row Filter node by choosing following options:



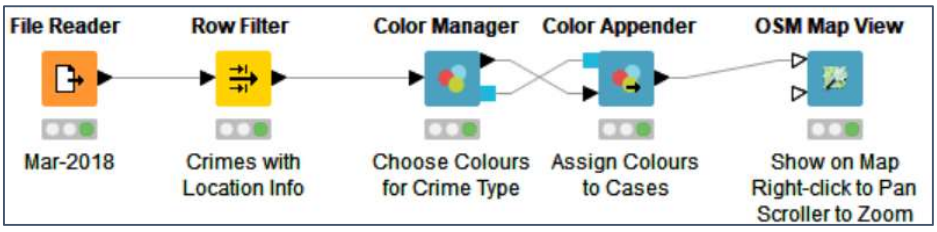



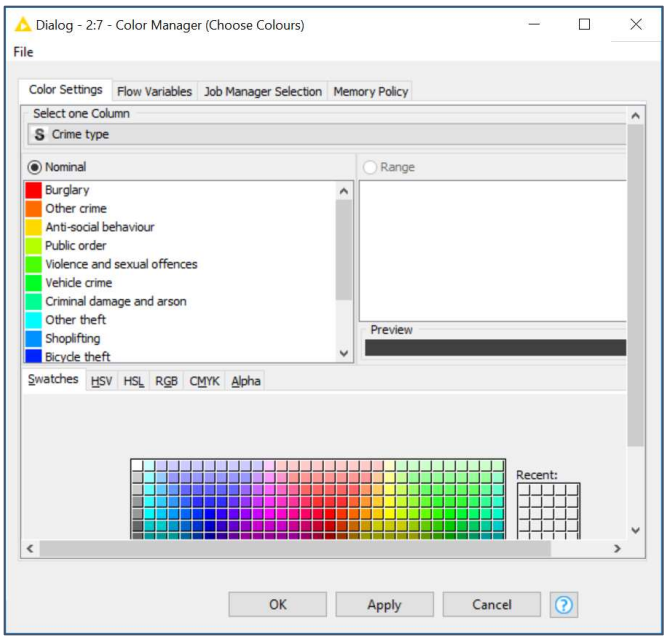
## 1.5 ACTIVITY 5 COLOURING DATA POINTS WITH LABELS

Now that we are familiar with the nodes and data flow, let's add some colours to our data so that we can see different types of crime cases using different colours.

►Search for two nodes called "Color Manager" and "Color Appender" and drag/drop them onto your workflow. Now connect them in your workflow as shown here:




►Now configure the Color Manager by using the column "Crime type" and feel free to edit the colours for each type of crime.




►After configuring the colour scheme, press OK to close this configuration and open "Color Appender" node to assign each case to their respective colours.

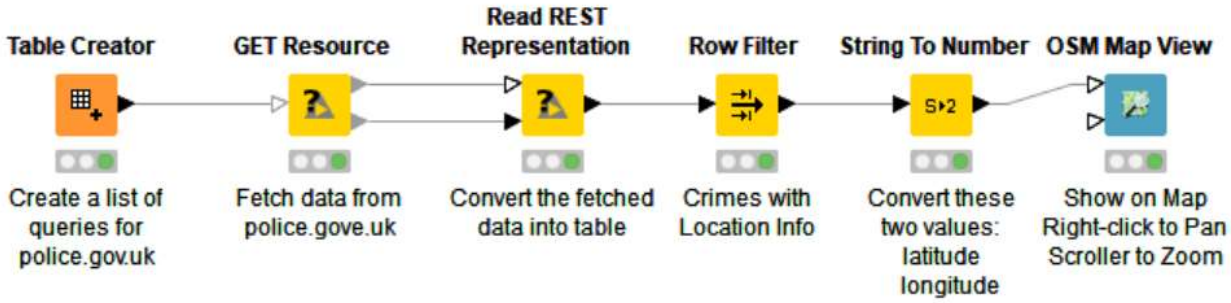
►That's it. Now rerun the workflow and open the OSM Map View. You can now see the coloured version of the crimes plot.

►You can see the data on Bing map as well, choose it from the top menu.

## 1.6 TAKE HOME ACTIVITY 6 FETCHING DATA FROM POLICE.GOV.UK IN REAL TIME

Now that you have practised some KNIME workflows, I suggest you to practise some on your own. Here is an example workflow for fetching Street Crime data from the UK Police website:

https://data.police.uk/api/crimes-street/all-crime?poly=52.268,0.543:52.794,0.238:52.130,0.478&date=2017-01
(Feel free to change the GPS coordinates in the polygon area or the date variable)




This workflow is for guidance purpose, I am quite sure that you can configure each of these nodes by reading their descriptions (or trial and error).

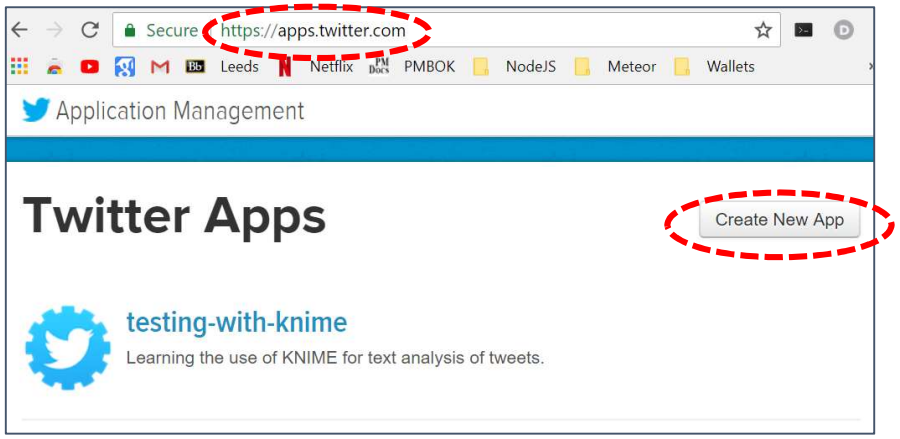
KNIME Archive for this **Workflow:** MyFirstWorkflowForCrimeAPI.zip
https://bit.ly/2rJlZzT

# 2 Twitter Text Analytics in KNIME

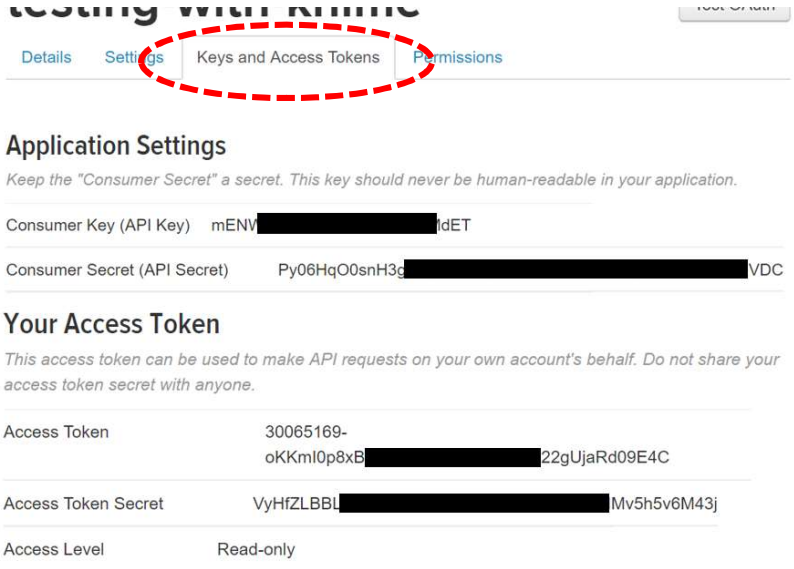
Our aim is to familiarise ourselves with text acquisition, pre-processing, and sentiment analysis in KNIME.

## 2.1 ACTIVITY 1 CREATING TWITTER APPLICATION FOR FETCHING TWEETS

►Visit this website: https://apps.twitter.com and login with your Twitter account. Create a New App with the button on top-right of this page. You will be asked few options, enter your preferences. Please note that the form asks for two web links, first one for your website, and second for a redirect link for oauth. You can provide your office or personal website in the first link. However, the redirect URL must be left empty (you will know this if you are a web or API developer).




►Now open the settings for newly created App, and visit the page titled "Keys and Access Tokens". You will get the API Key and API Secret from Application Settings.




►Also, generate a new Access Token that will be used to fetch data from KNIME.

If you have no Twitter account yet, you can create one or if you don't wish to create one, you may use a temporary access token that I have created

| | |
|---|---|
| Consumer Key (API Key): | ocURT3yZxUzAhpCA3MuvxK3AL |
| Consumer Secret (API Secret): | QrVKkHmPH3TYHdC5Yd06tg8DXhUuAWo0YmllcmhqyXL6v0x7MD |
| Access Token: | 30065169-MeZRCgVmizww4fFSOJCTxOkMJ9QQXVqGVrOGeQC6z |
| Access Token Secret: | Rq9hVakAIvyvnzpfc4NOCnuCsOIG0FIuxGJT73BioHAlF |

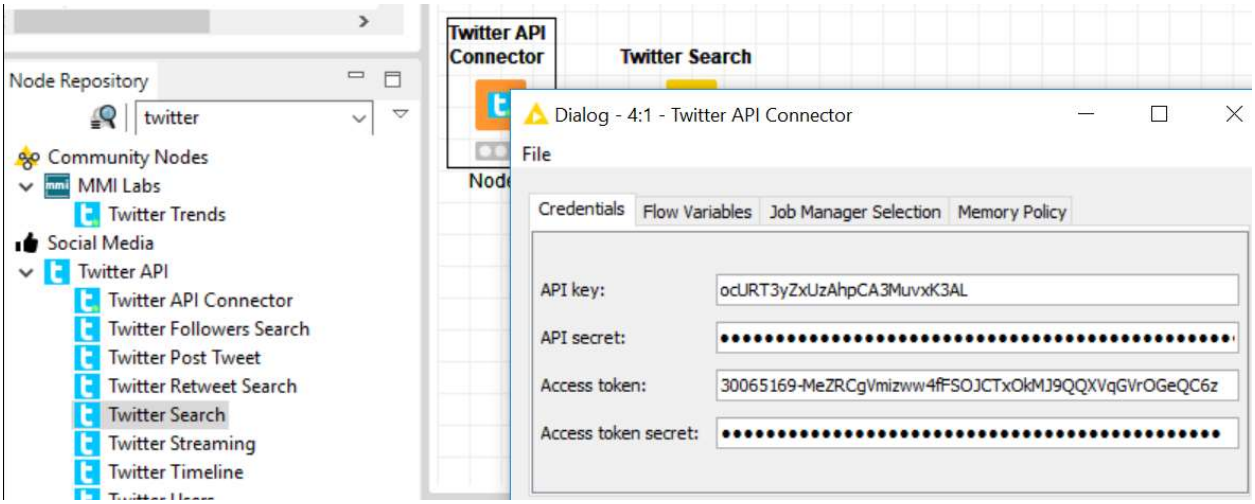
These keys will be removed after this workshop, and hence not for future use.

Now we are ready to use Twitter API through KNIME, so let's begin.

## 2.2 ACTIVITY 2 ACCESSING TWEETS FROM KNIME

Create a new workflow by choosing "New…" from the File menu on the top-left of screen.

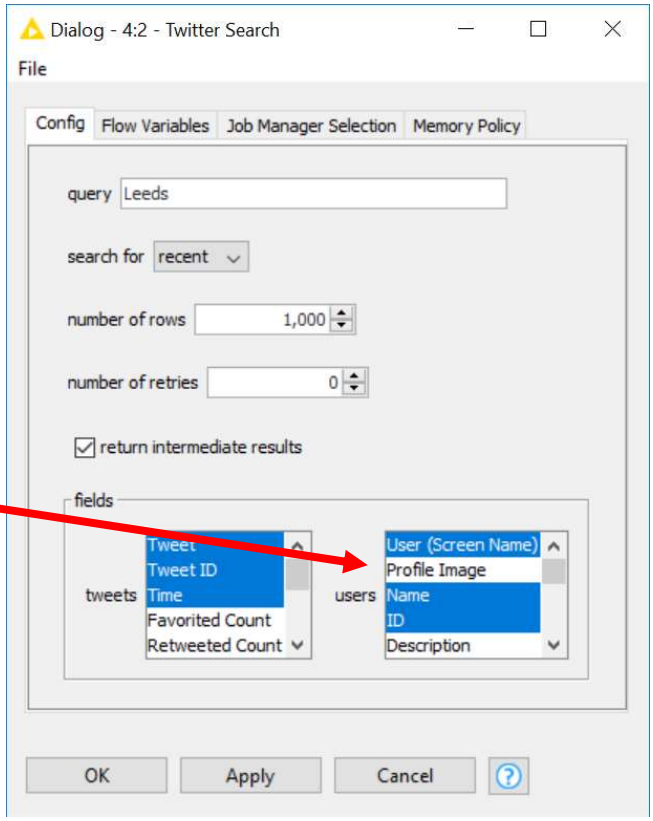
►**Select "New KNIME Workflow" and click next.** Feel free to choose anything that you find appropriate for example **"MyFirstTwitterAnalysis".**
►Search "Twitter" in the Node Repository window, and drag/drop the two nodes "**Twitter API Connector**" and "**Twitter Search**" on your workflow.
►Double-click to configure **Twitter API Connector** with the information that we generated in the last activity (see last page). Execute it to confirm that your connection is successful.




►Now we can search for tweets using the "**Twitter Search**" node.

Let us start with simple search, we will query 5,000 (or 1000 if your system is slow) most recent tweets containing the word "Leeds".

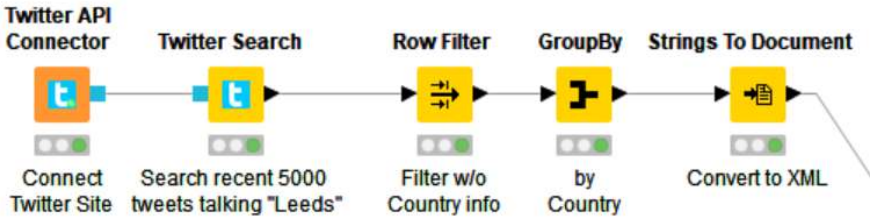
It is important to note that fetching Profile Image for each and every tweet is not a good idea because it will unnecessarily choke the internet speed as well as processing. Deselect this information from the configuration box as shown here




►As you should be feeling comfortable with KNIME user interface by now. I will provide less details on each step from this point onward.

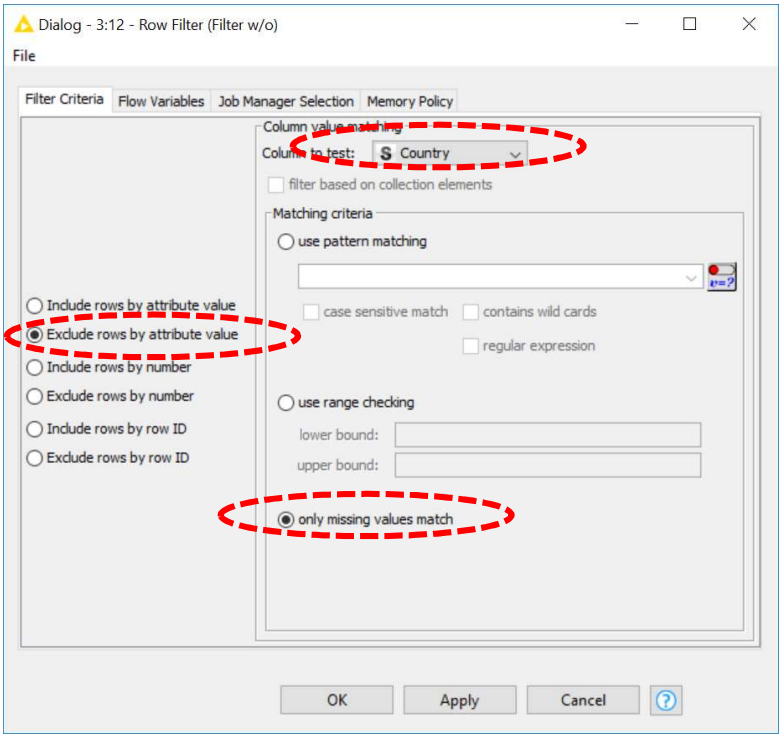
## 2.3 ACTIVITY 3 GROUPING ALL TWEETS COUNTRY-WISE

►We can perform several different types of analyses with these tweets. Let's do a simple one to begin with. We will group all tweets according to countries and will prepare tag cloud for most common terms found in each set of tweets.

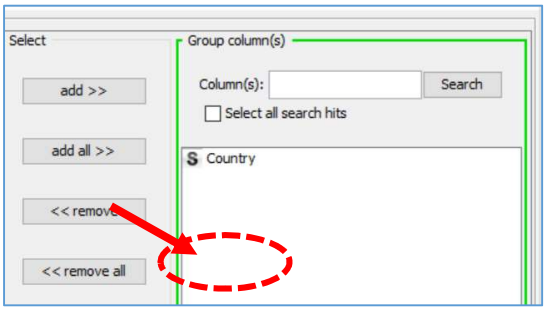


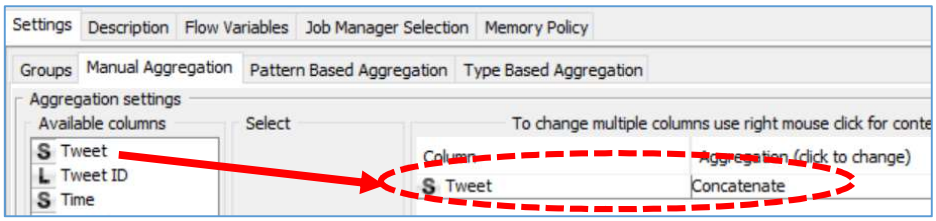

►Use the following screenshot to expand your workflow with **Row Filter**, **GroupBy**, and **"Strings To Document"** node. We will configure these nodes one by one.

►The Row Filter node is required to drop all those tweets where we don't know which country they emerged from. Follow this screenshot to configure it properly:




►The GroupBy node is used to combine all the tweets coming from same country. Configure two pages "Groups" and "Manual Aggregation" for this node, as shown below.



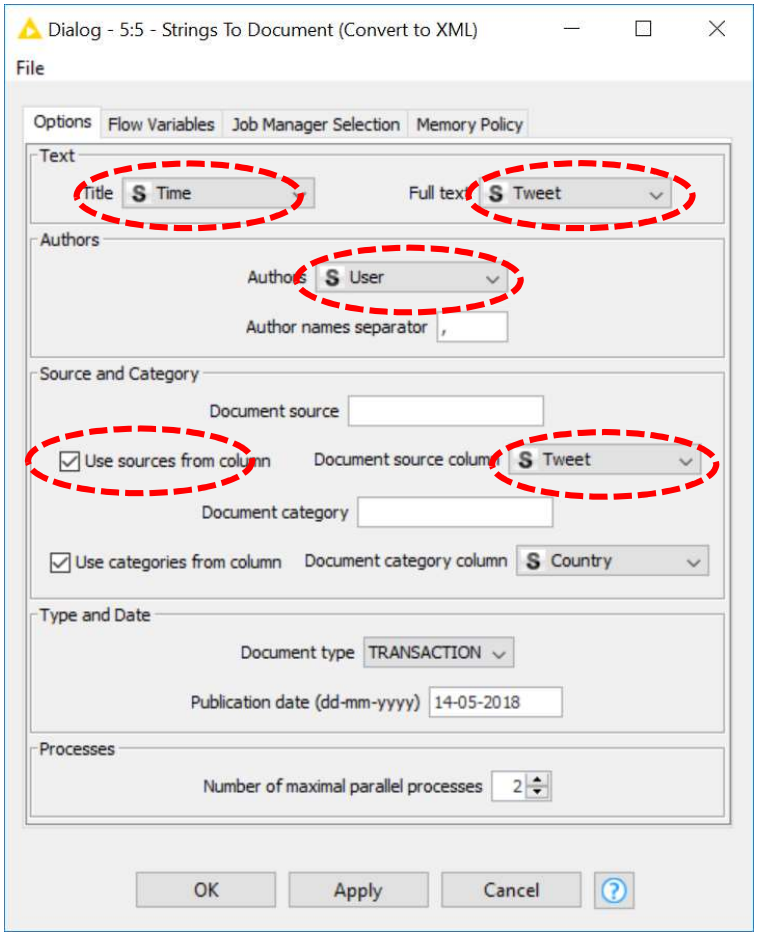



## 2.4 ACTIVITY 4 CONVERTING TWEETS INTO DOCUMENT FOR ENRICHMENT

►We need to convert the strings of textual data into XML document using "Strings to Document" node. This is an important step as XML documents can be enriched with additional tags and modifiers without losing the original contents.

*Thereafter, we usually apply number of filters that will remove numbers, small words, punctuations and stop words. All the words are converted to lower case (or upper case) so that words appearing at the beginning of a sentence should not be treated different from same word appearing somewhere else. Then we perform stemming and lemmatization, which is another important pre-processing step. To understand its important, consider the example: go, goes, and going. These words can all be combined as just one concept: "go".*

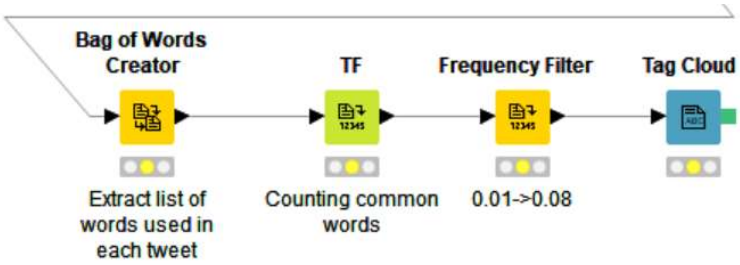
*We are skipping all these operations in this tutorial to keep it simple.*




## 2.5 ACTIVITY 5 FINDING FREQUENCY OF EACH TERM AND CREATING TAG-CLOUD

►We can now convert the plain descriptive text into a list of words and count the frequency of each term inside the documents. Find and place the following four nodes in your workflow, and configure them yourself.

Please note that you can read each node's description on the top-right frame in KNIME User Interface.

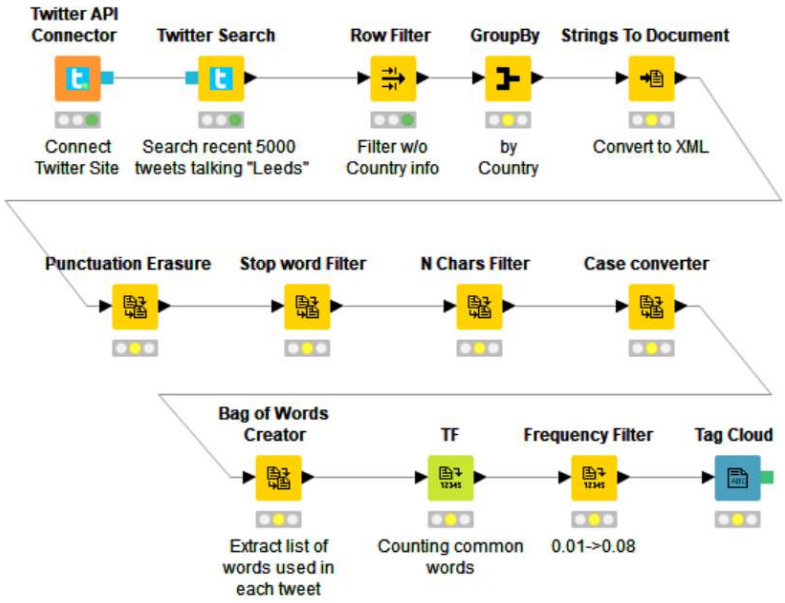



►For the Frequency Filter, I chose 0.01 to 0.08, but feel free to play with this.

## 2.6 TAKE HOME ACTIVITY 6 REFINING THE TAG CLOUD

Now that you have practised some Twitter search and text analysis, I suggest you to practise some on your own.

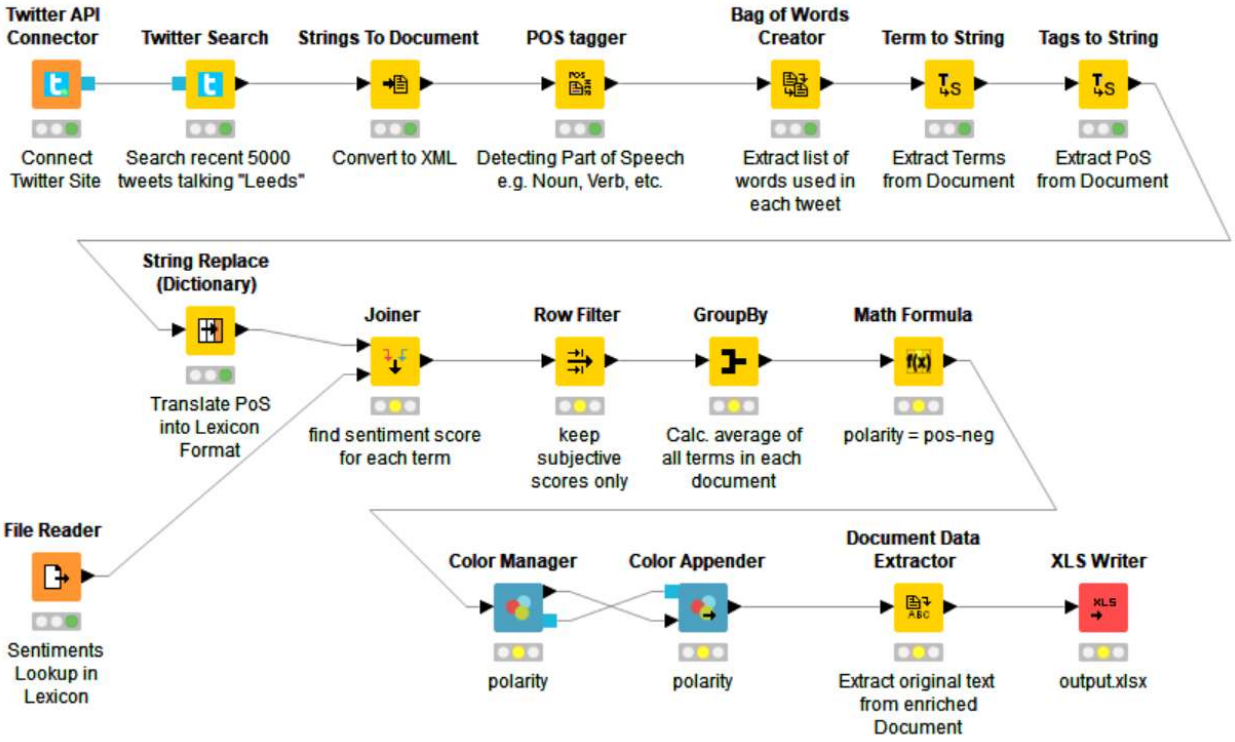
Here is an example workflow for refining the tag cloud using a series of filters like removing punctuations, stop words, small words, and changing all cases to lower case (or upper case).




KNIME Archive for this **Workflow:** MyFirstTwitterWorkflow.zip
https://bit.ly/2L4hGY6

# 3 Sentiments Analysis of Tweets

Our aim is to familiarise ourselves with sentiments analysis in KNIME. Create the following workflow diagram, and configure each node correctly with the help of node description (and trial and error, if you wish).




For this workshop, we can save time by importing workflow instead of creating it from scratch.

Please download these files:

KNIME Archive for this **Workflow:** MySecondTwitterWorkflow.zip
https://bit.ly/2IptuCj

Sentiments Lookup file for **File Reader:** sentiments.csv
https://bit.ly/2wJF6yH

PoS Dictionary for **String Replace:** dictionary-partsofspeech.txt
https://bit.ly/2IDEC1U