Hands-on Tutorial on Topic Modeling using KNIME

Raghava Mukkamala (rrm.digi@cbs.dk)

This hands-on exercise will use the KNIME Analytics Platform to build a topic modeling KNIME model on the textual data from the IMDB online reviews dataset.

Prerequisites

Before starting this exercise, you should have installed the KNIME Analytics Platform on your computer.

If you have not installed the KNIME Analytics Platform on your computer, then please go through the following URL to download. https://www.knime.com/downloads.

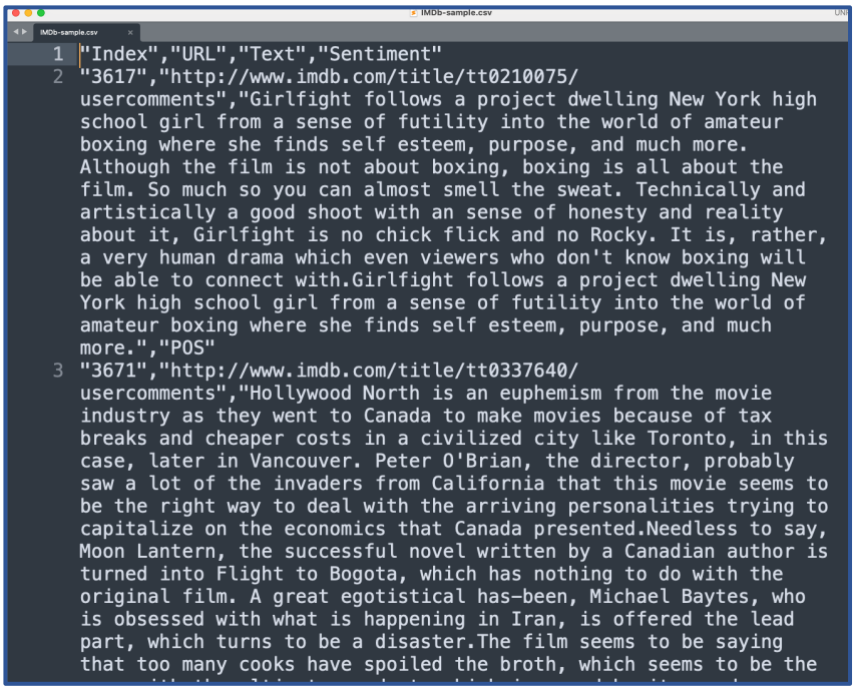
Here is the installation guide to KNIME:
https://docs.knime.com/latest/analytics_platform_installation_guide/index.html#_installing_knime_analytics_platform

1 Dataset: Online movie reviews (IMDb-sample.csv)

Before we begin creating a KNIME workflow to build a Topic model, we must understand the dataset first. The online movie review dataset is provided to you as part of this hands-on session in the form of a CSV file. Therefore, we will first explore the attributes/column names of the dataset displayed in the following table.

Data Attribute	**Explanation**
Index	Review Index
URL	Old link to the movie review (ignore the link)
Text	Review text
Sentiment	Sentiment Label: POS, NEG. (**dependent variable or target label**)

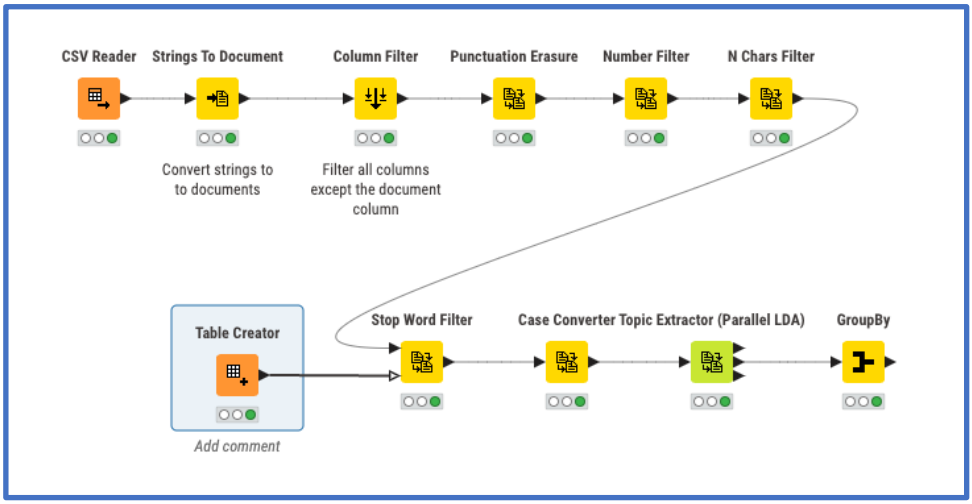
The first few records in the dataset are shown below.



Index *Number (integer)*	URL *String*	Text *String*	Sentiment *String*
6168	http://www.imdb.com/title/tt0217...	Hmm, IMDb rating of 7.5, good comments, bla, bla ... okay, two of my friends and me, we ord...	NEG
3963	http://www.imdb.com/title/tt0433...	How many fricken' times do we have to see a spook walking by in the background & peaking...	NEG
7173	http://www.imdb.com/title/tt0113...	I HATE plane crash movies...ALL of them! In fact, I hate them all with a passion! First of all, t...	NEG
5118	http://www.imdb.com/title/tt0805...	I am a Christian and I say this movie had terrible acting, unreal situations and a completely f...	NEG
8693	http://www.imdb.com/title/tt0840...	I am from the Dallas/Fort Worth area and lived in Arlington for a few years. This movie was ...	NEG
12305	http://www.imdb.com/title/tt0494...	I am watching this movie right now on WTN because that was the channel that the TV was t...	NEG
4438	http://www.imdb.com/title/tt0253...	I borrowed this movie from library think it might be delightful. How wrong am I!It is such a b...	NEG
6091	http://www.imdb.com/title/tt0273...	I bought this movie exciting a gloriously gratuitous, over the top, entertaining bloodbath. I go...	NEG
6117	http://www.imdb.com/title/tt0083...	I can't say this is the worst movie ever made, but personally I think of it that way because wh...	NEG
4198	http://www.imdb.com/title/tt0309...	I caught this Cuban film at at an arthouse film club. It was shown shortly after the magisteri...	NEG
30	http://www.imdb.com/title/tt0127...	I did not like the idea of the female turtle at all since 1987 we knew the TMNT to be four brot...	NEG
2820	http://www.imdb.com/title/tt0108...	I don't know much about Tobe Hooper, or why he gets his name in the title, but maybe he sh...	NEG
8129	http://www.imdb.com/title/tt0071...	I dunno sometimes...you try and try and try to be charitable towards all the B thru Z grade m...	NEG
3118	http://www.imdb.com/title/tt0448...	I felt that the movie was dry... very disappointing no plot..kept waited for something to happ...	NEG
12277	http://www.imdb.com/title/tt0074...	I found The Arab Conspiracy in a bargain bin and thought I'd uncovered a lost treasure. Folks...	NEG
9192	http://www.imdb.com/title/tt0317...	I found this movie in the 'horror' section of my video store. That seems to make sense as m...	NEG
7270	http://www.imdb.com/title/tt0783...	I gave this looooooooooooong film a 2 because of the attractive actors and semi-sexy love sc...	NEG
10283	http://www.imdb.com/title/tt0480...	I hate to sound like an 'old person', but frankly I haven't seen too many movies that I like that ...	NEG
8668	http://www.imdb.com/title/tt0086...	I have never read the Bradbury novel that this movie is based on but from what I've gathered,...	NEG
11689	http://www.imdb.com/title/tt0454...	I have never seen any of Spike Lee's prior films, as their trailers never caught my interest. I h...	NEG

2 Build KNIME workflow for Topic Modeling

After completing the hands-on exercise, the final KNIME workflow will appear, as shown in the following figure.



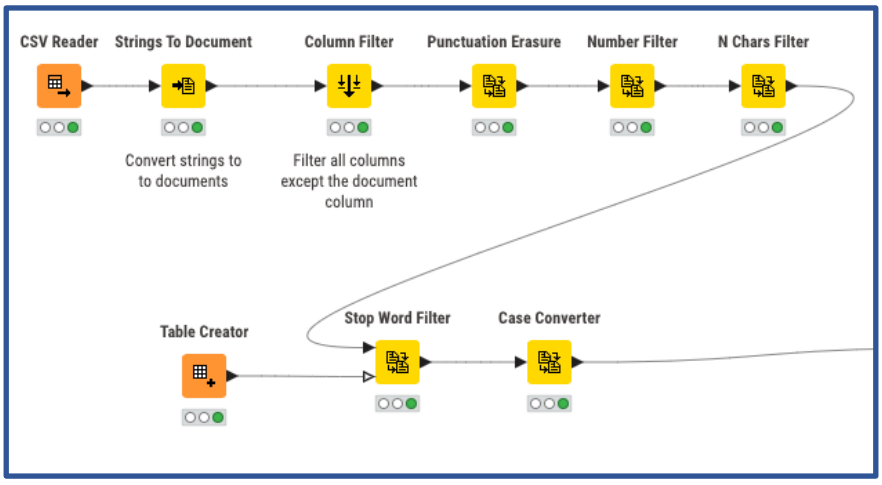
You will build a KNIME workflow to build a topic modeling model using the Online reviews dataset. The process involves the following steps.

1. Reading and pre-processing the data
2. Build Topic Models

We will go through each of these steps in more detail below. Before you start, if you have not watched previously, we recommend that you watch the video (only 2 minutes duration) named “What is a Node? What is a Workflow?” (https://www.youtube.com/watch?v=4rETNe-Xx7k) one more time to recap the basics of how to build and execute KNIME Workflows.

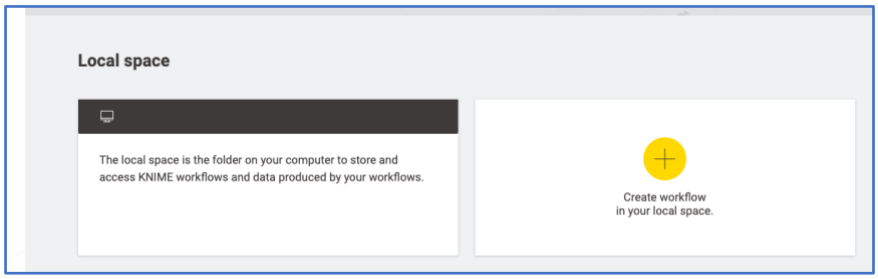
2.1 Reading and pre-processing the data

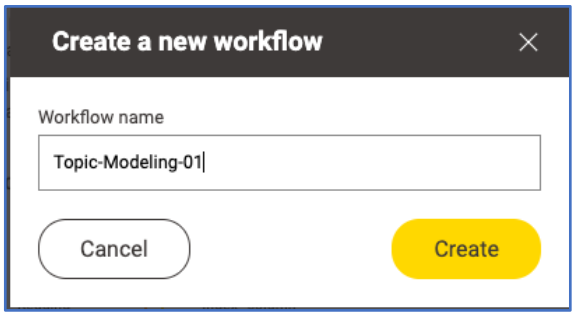
This step involves two nodes, as shown in the following picture.

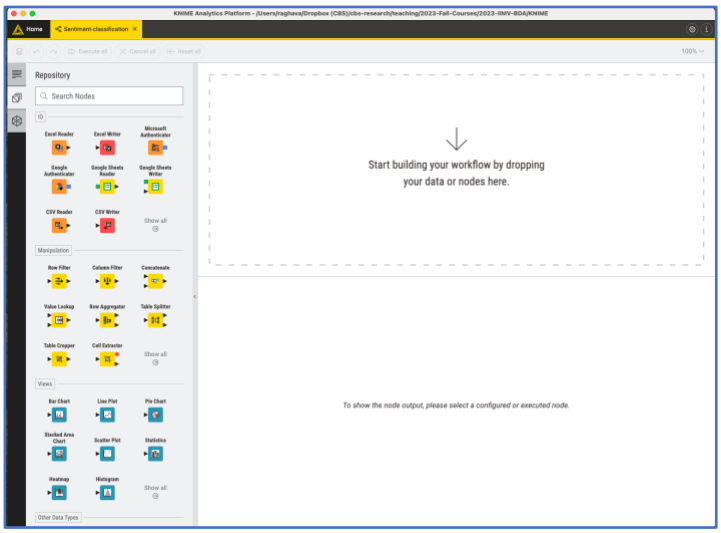


First, you should download the online movie review dataset (*IMDb-sample.csv*) that is supplied to you as part of Datasets in this hands-on session. Download the dataset to your computer (if you have not done that already) and make sure that you know the exact location of the file on your computer so that you can later open the file in KNIME. On Windows, you use the *File Explorer* program to browse folders and files. On Mac, you can use *Finder*.

When you have downloaded the dataset file, open the KNIME Analytics Platform, right-click your local workspace in the KNIME Explorer, and select New KNIME Workflow…. Select a name for the workflow and click Finish as shown in the following screenshots.



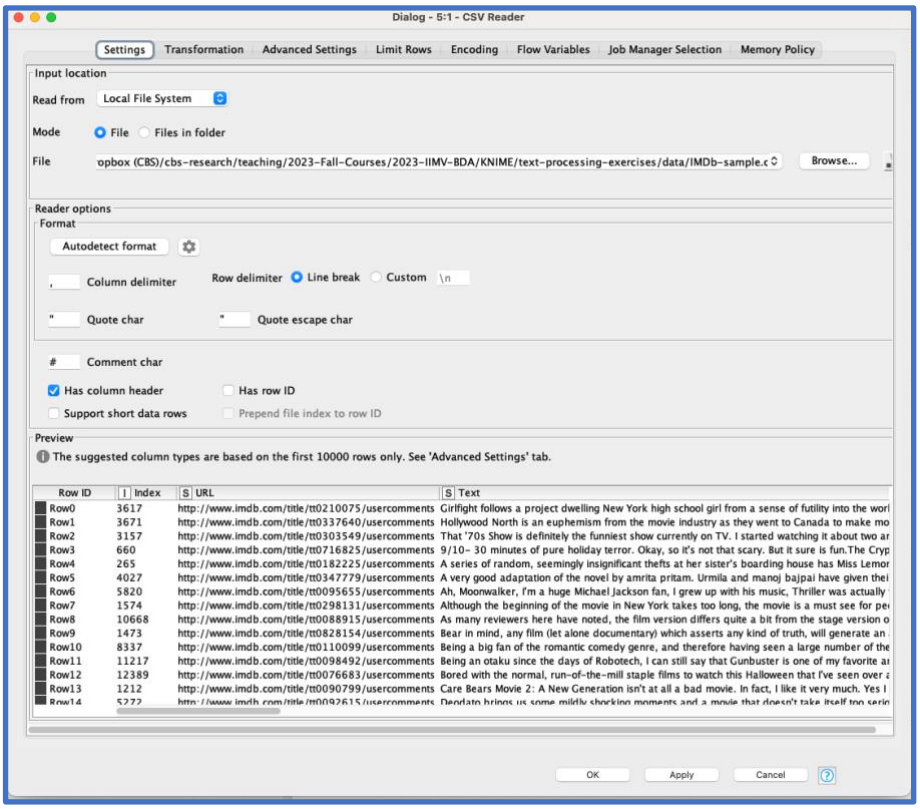




2.1.1 CSV Reader Node

Since the dataset is an excel file, we will use an ***CSV Reader*** node to read the data. You can find an ***CSV Reader*** in the Repository by typing "***CSV*** Reader" in the search field.

Drag an ***CSV Reader*** node to the Workflow Editor. Then double-click the ***CSV Reader*** to open its configuration dialog. In the configuration dialog, click the "browse…" button and try to find the .csv file (*IMDb-sample.csv*) containing the dataset. Please make sure to configure it as shown below.



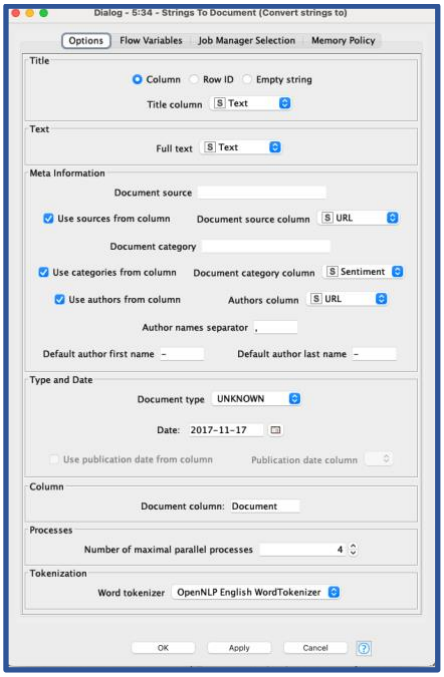
If you cannot find the dataset file from the File Reader configuration dialog, use the following approach instead:

1. Open File Explorer if you are using Windows or Finder if you are using Mac.
2. Open the Downloads folder or the folder where you have downloaded the files.
3. This folder should contain the dataset file you downloaded from the course page.
4. Drag the dataset file and drop it on the KNIME Workflow Editor. This will automatically create a File Reader that is configured to read the dataset file that you dragged and dropped on the Workflow Editor.
5. Make sure that the settings are assigned as shown in the above figure.

When you have configured the csv Reader, then right-click it and select Execute. When you click the node, it will show data from the Excel file in the File Table below as follows.

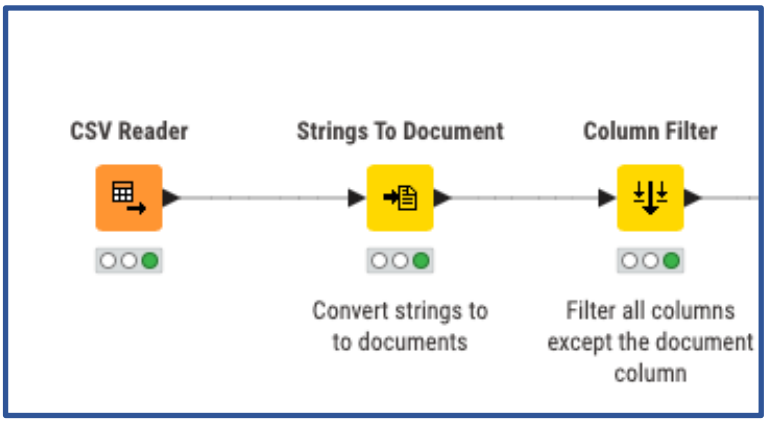
2.1.2 String to Document Node:

Search for String to Document node and connect to the input from the CSV reader. Choose the *title* column and *Full text* column as text as shown below. The rest you can leave as it is.

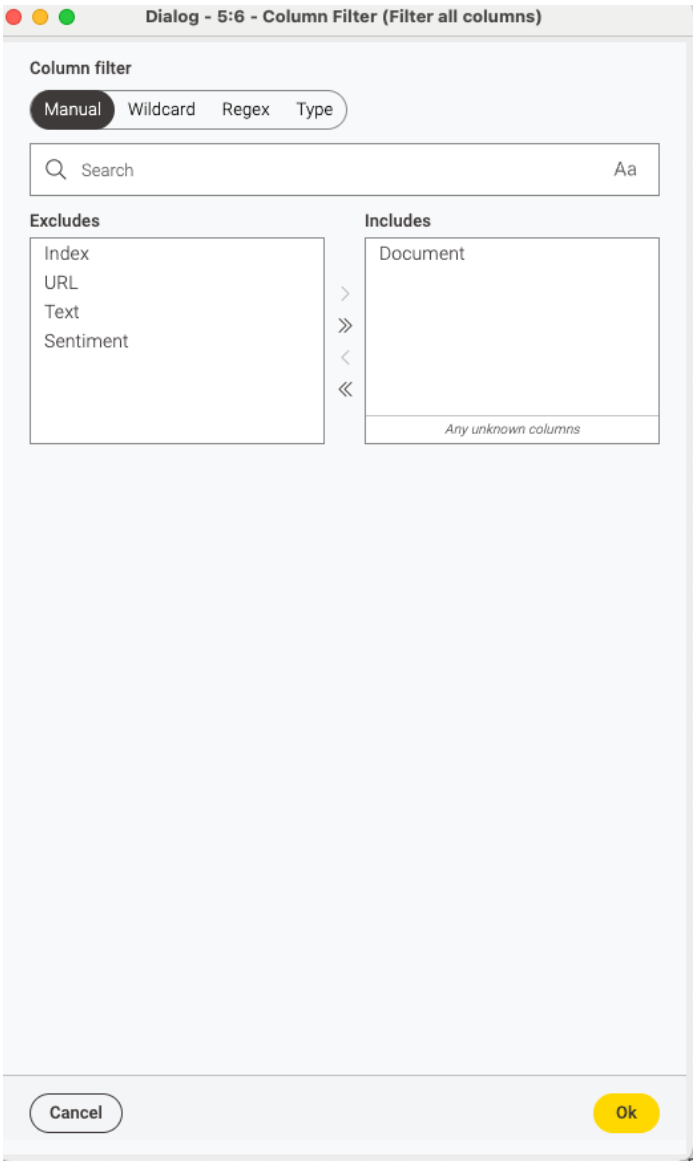


2.1.3 Column Filter Node

Search for the column filter node and connect it, as shown below.

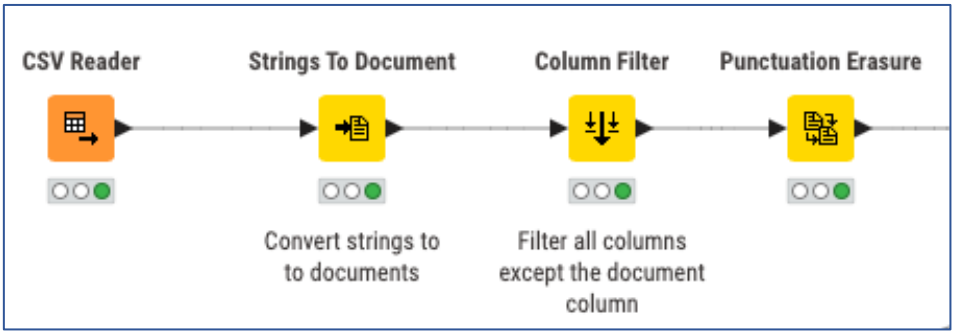


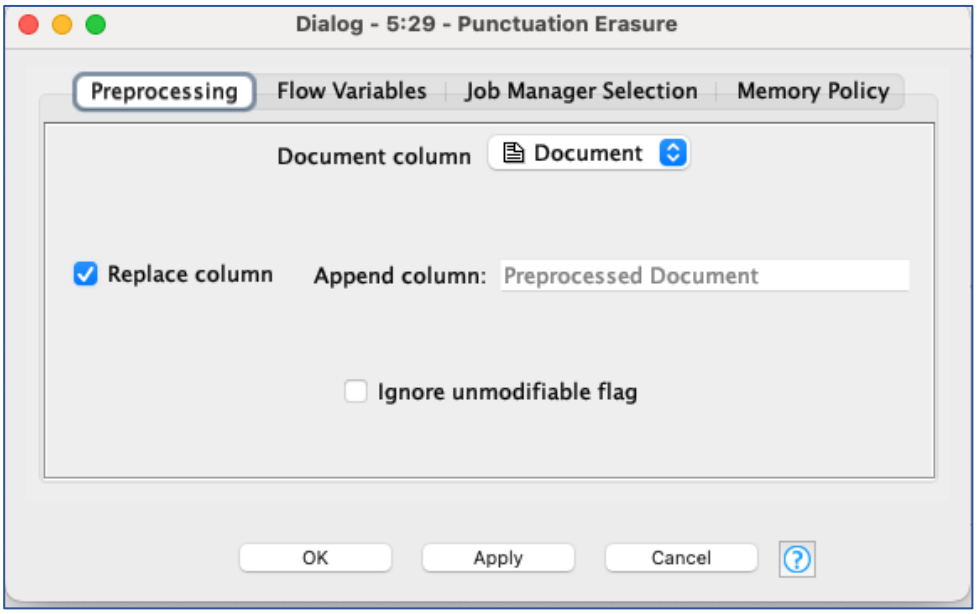
In the settings for the Column Filter node, exclude all other columns except Document, as shown below.



2.1.4 Punctuation Eraser Node

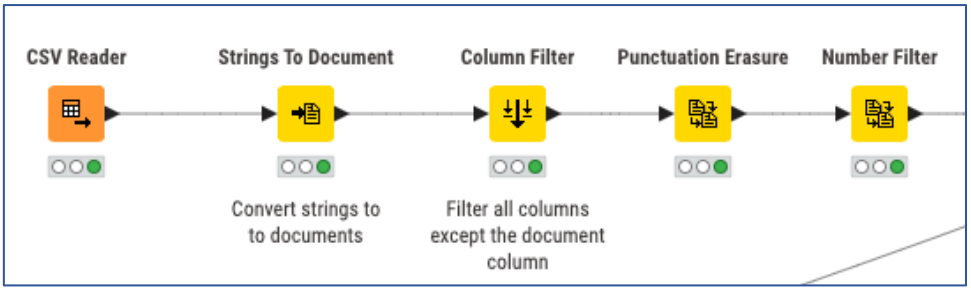
Search for the Punctuation Eraser node, connect it, and configure it as shown below.

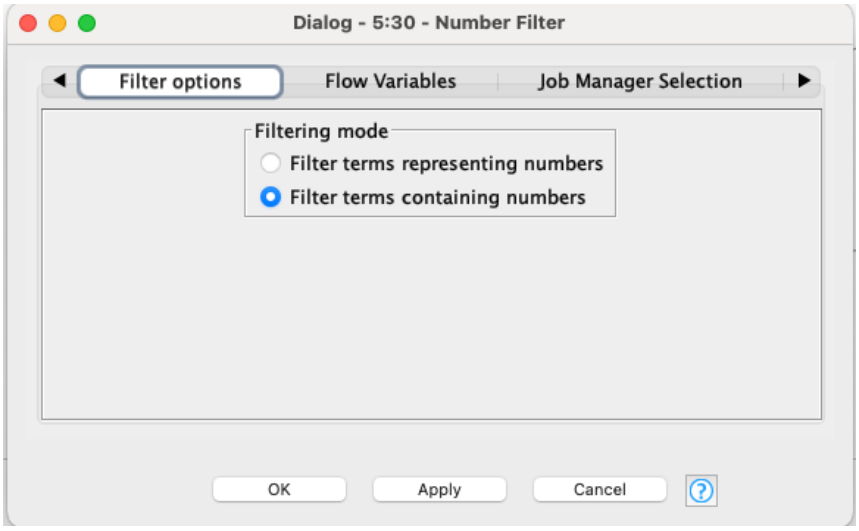




2.1.5 Number Filter Node

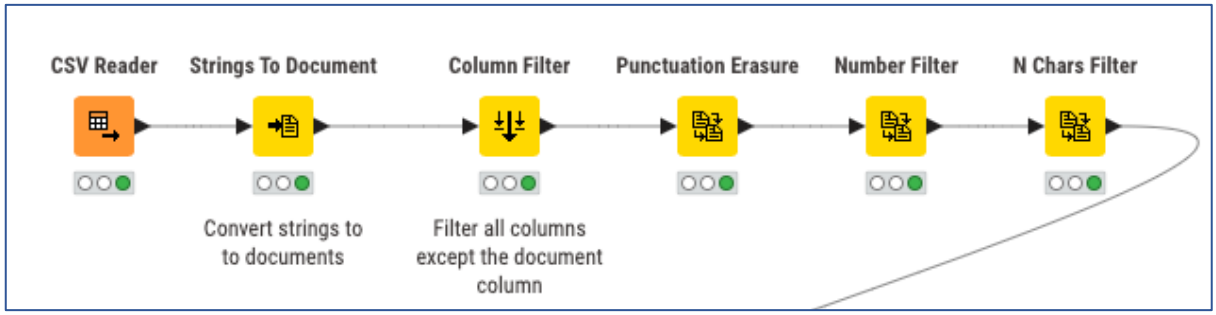
Search for the Number Filter node to filter numbers from the text, connect it, and configure it, as shown below.

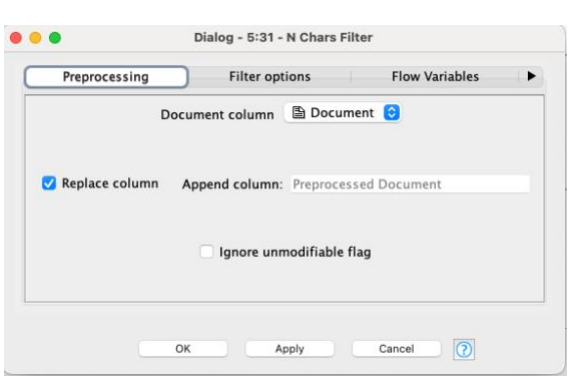


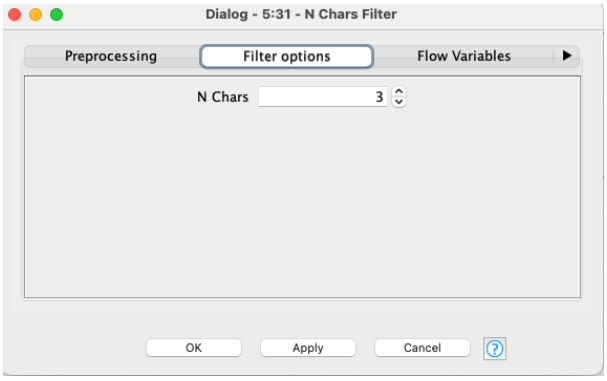


2.1.6 N Char Filter Node

Search for the N Char Filter node, connect it, and configure it, as shown below. It will remove the words with less than the specified number of characters.

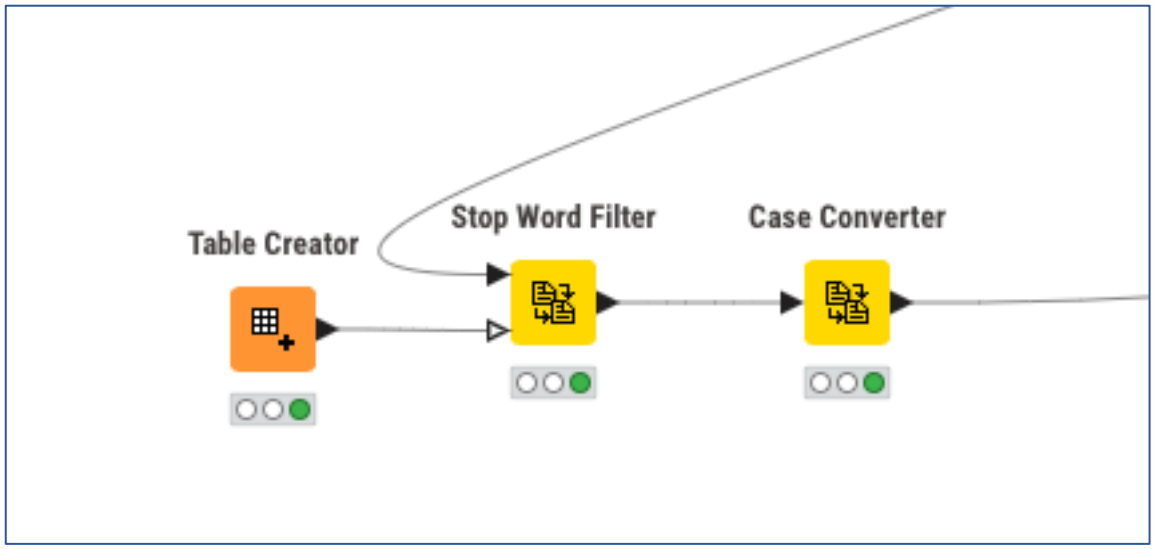




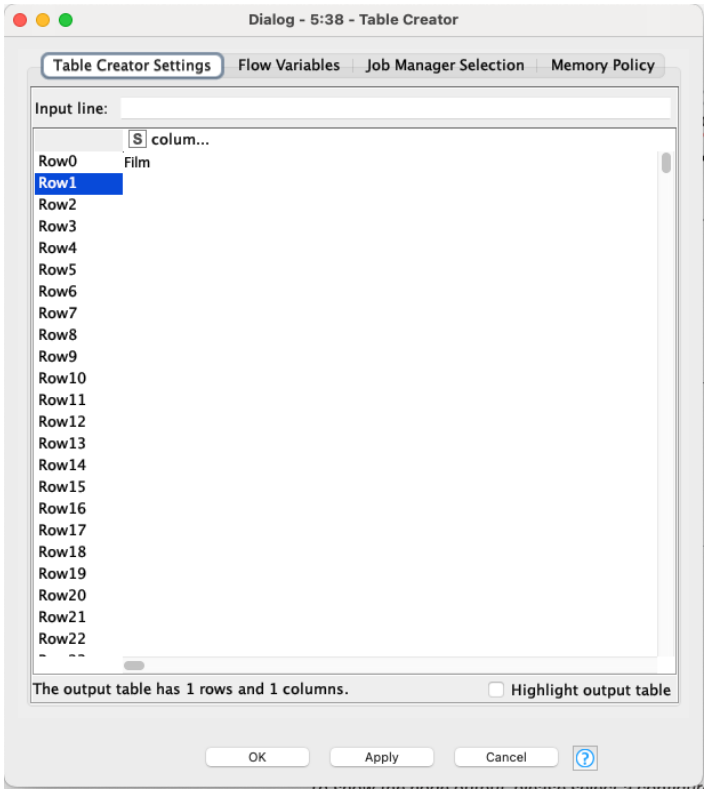


2.1.7 Table Creator Node (for custom stop words)

Search for the Table creator node to create the custom stop words as shown below slides.

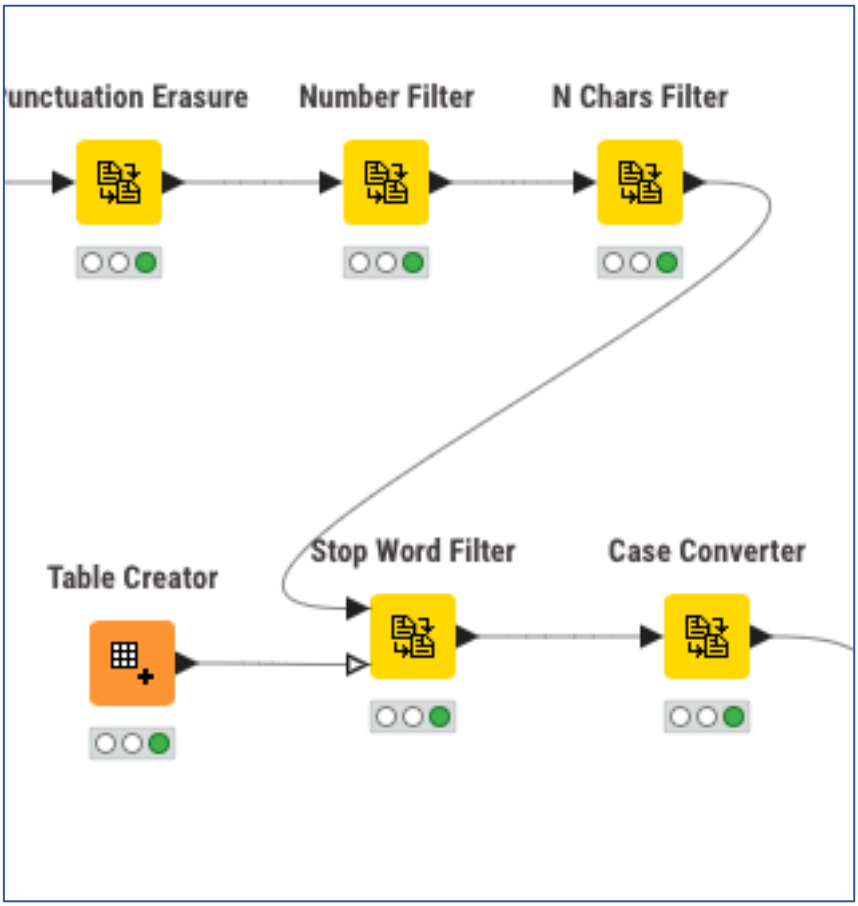


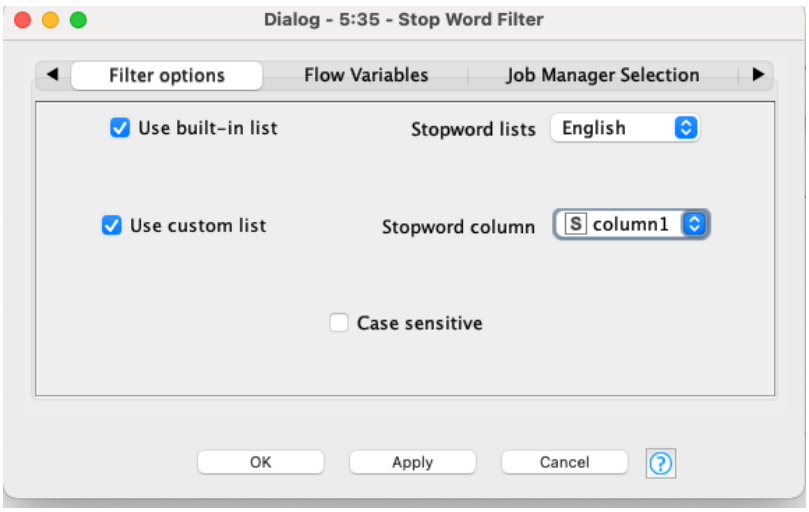
In the configuration of the node, you can add whatever custom stopwords we want to remove from the reviews. In the reviews, normally, we will have the word Film very reportedly, so we will remove it as a custom stop word.



2.1.8 Stop Word Filter node

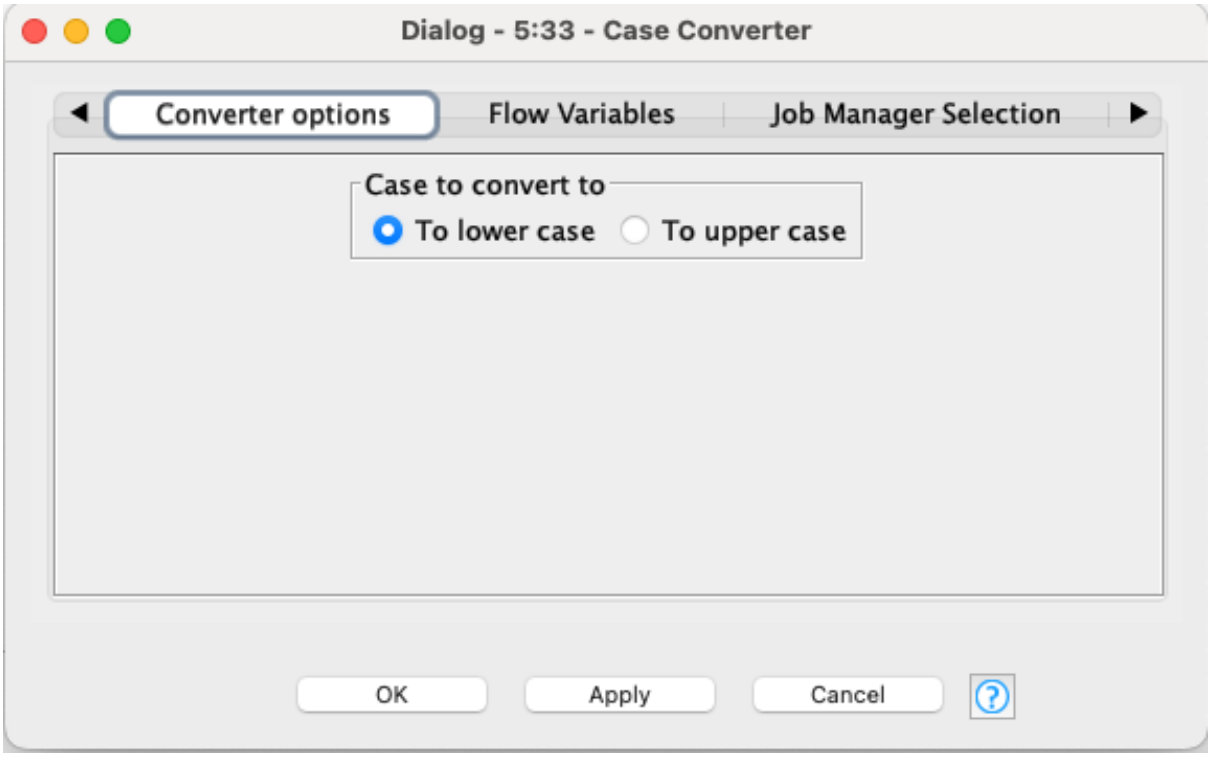
Search for Stop Word Filter Node and connect the node to other nodes as follows.



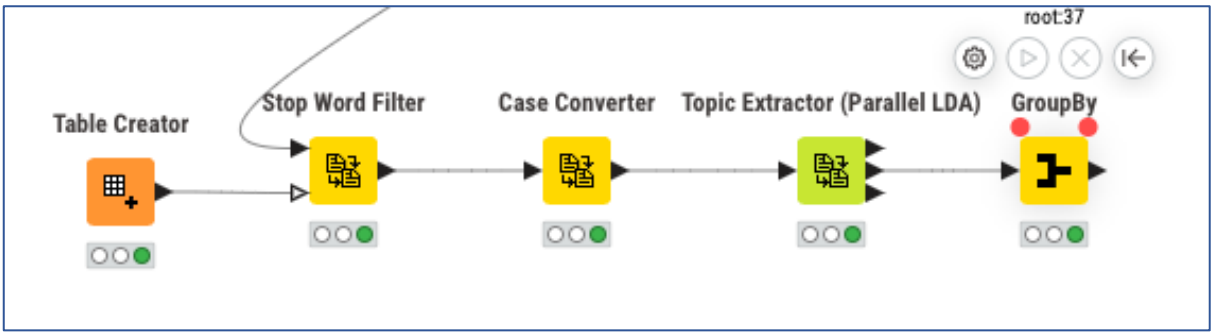


2.1.9 Case Converter node

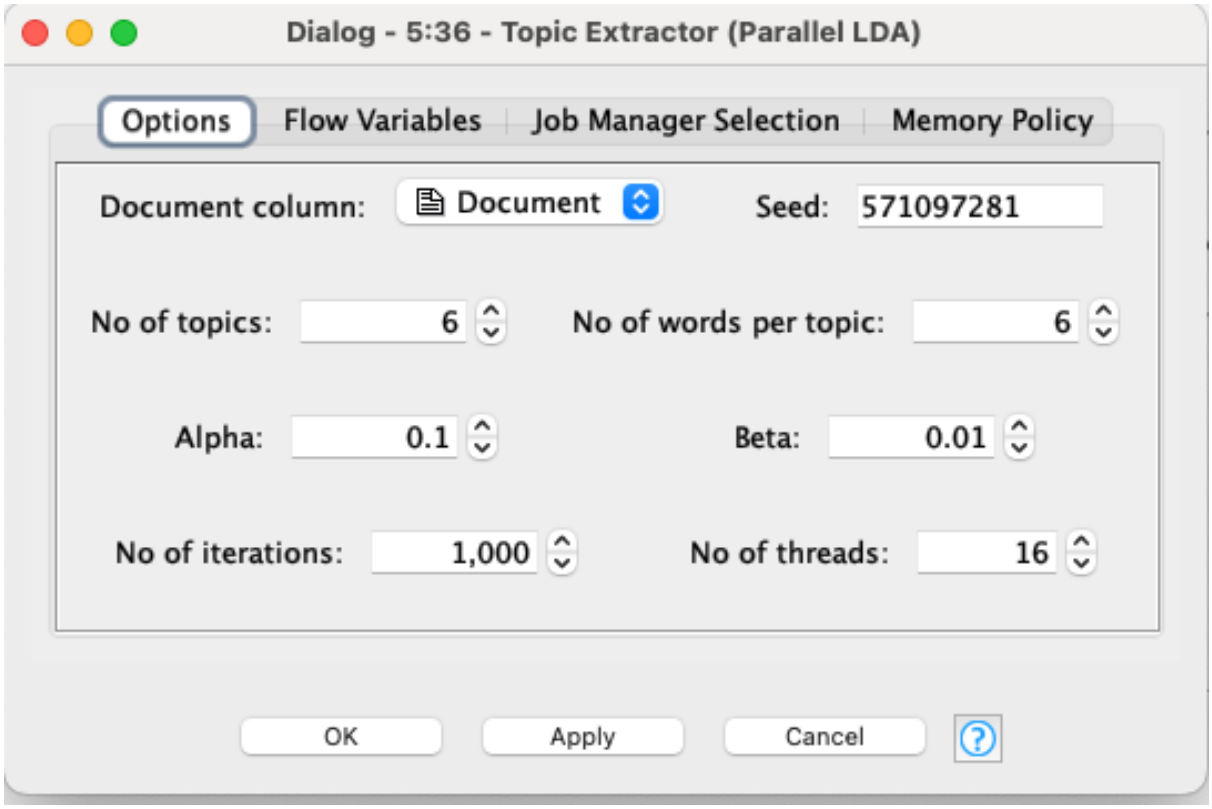
Search for the Case Converter node and configure it as follows. We want to convert the document to lowercase, so choose To Lower Case.



2.1.10 Topic Extractor node



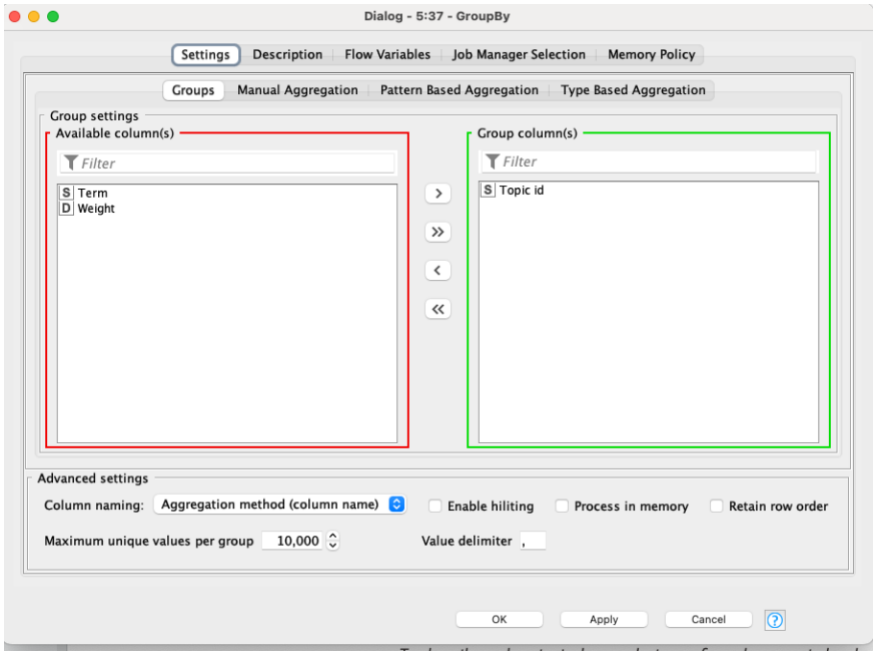
Connect the Topic Extractor node as shown above and configure the node as shown below. No need to change any settings and use the defaults.



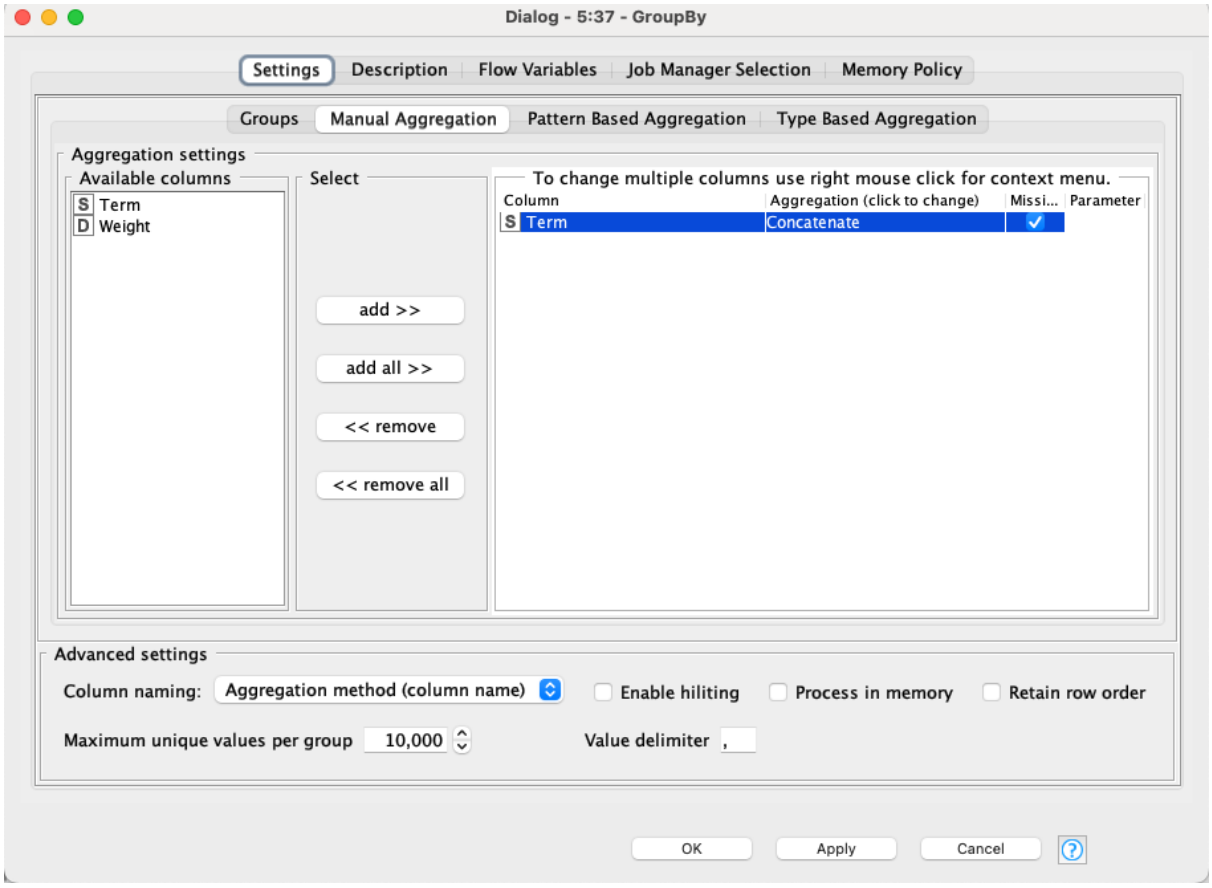
2.1.11 Group By node

Search for Group By node and connect as shown in the screenshot above.

Click Groups and include the topic Id as shown below.



Click the Manual Aggregation tab and configure the settings as follows.



Finally when you all the nodes, you can see the topics and their words as shown below. Note that Topic extractor node will take some time to compute the topics.

#	Row...	Topic id (String)	Concatenate(Term) (String)
1	Row0	topic_0	people, time, war, movie, world, documentary
2	Row1	topic_1	movie, bad, time, acting, people, horror
3	Row2	topic_2	story, character, time, little, movie, life
4	Row3	topic_3	movie, time, story, love, people, movies
5	Row4	topic_4	little, movie, woman, time, character, love
6	Row5	topic_5	story, films, character, series, characters, original