Shantanu S. Mangale
BE B 41242 DMW

Assignment No. 4

Title: Consider a suitable text dataset. Remove stop words, apply stemming & feature & feature selection techniques to represent documents as vectors. Classify documents & elevate precision, recall.

Problem Definition:-
Remove stop words.

Pre-requisite:-
Basic concepts of ETL.

S/W & H/W Requirements:-
Rapid Miner, PIV, 2GB RAM, 500 GB HDD, Lenovo A13-4089 Model.

Learning Objectives:-
We are going to learn how to tokenize & filter a document into its different words & then do words count for each word in a text document.

Outcomes:-
You are able to see a word list containing all the different words in your document & their occurence count next to it in the

"Total Occurences" column.

Theory Concepts:-

Text Processing Tutorial with RapidMiner:

In this manual, we are gain to learn how to tokenize & filter a doument into its differentiate words & then do words count for each word in a text document.

Open RapidMiner & click 'New Process'. On the left hand panel of your screen, there should be a tab that says 'Operators' - this is where you can search & find all the operators for Rapid Miner & its extensions. By searching the Operators tab for 'read', you should get an output like this. There are multiple read operators depending on which file you have & most of them work the same way. If you have & most of them work the same way. If you scroll down, there is a 'Read Document' operator. Select this operator & enter it into your Main Process window by dragging it. When you select the Read Documents operator in the Main Process Window, you should see a file uploader in the right.

After you have chosen your file, make sure that the output port on the Read Documents operator is connected to the 'res' node in your Main process. Click

the 'play' button to check that your Main process. Click the 'play' button to check that your file has been received correctly. Switch to the results perspective by clicking the icon that looks like a display chart above the 'Process' tab+the top of the Main Process panel.

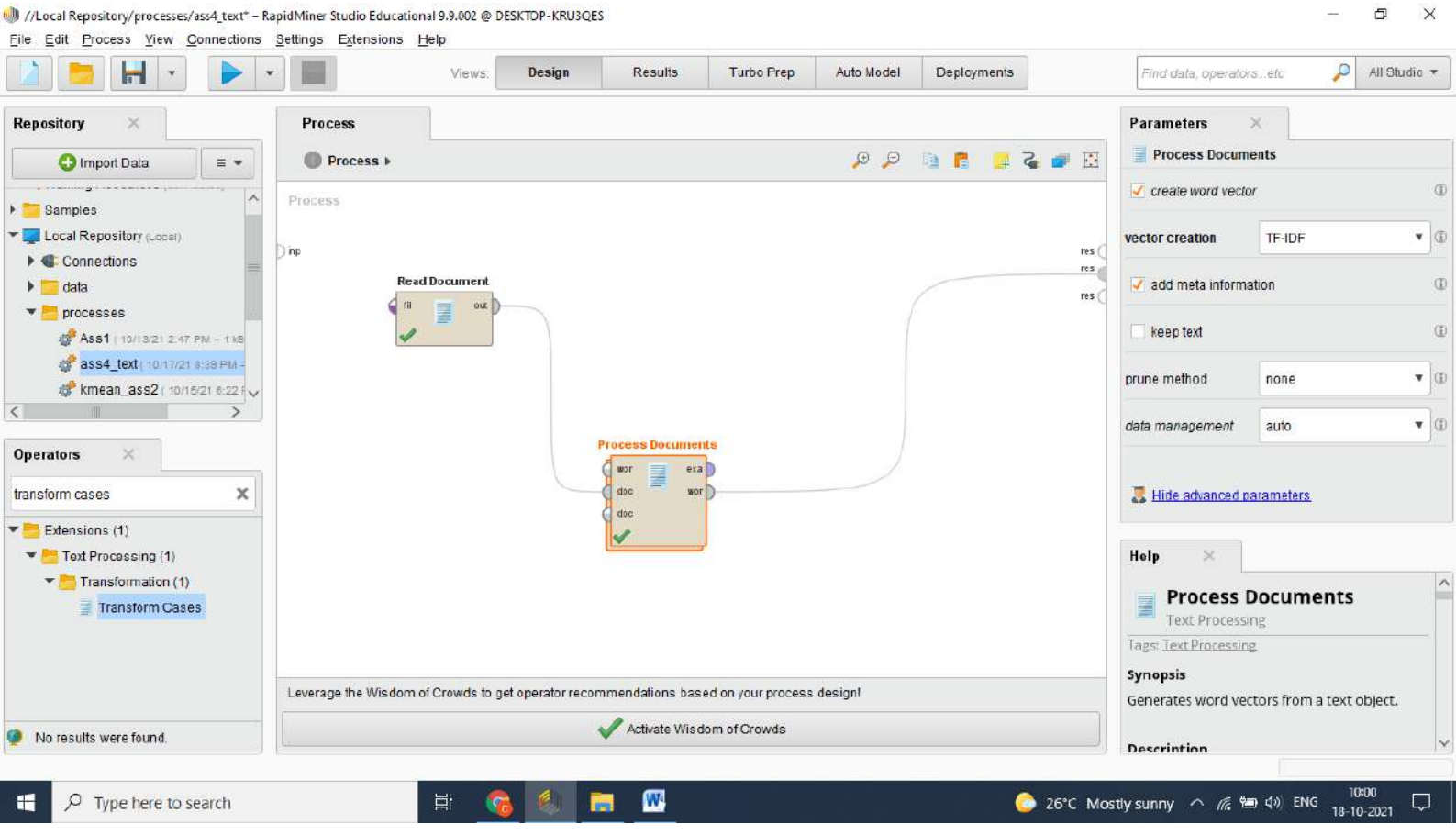
Now we will more on processing the ~~dom~~ document to get a list of its different words & their individual count. Search the Operator list for 'Process Document to get a list of its different words & their individual count. Search the Operator list for 'Process Documents'.

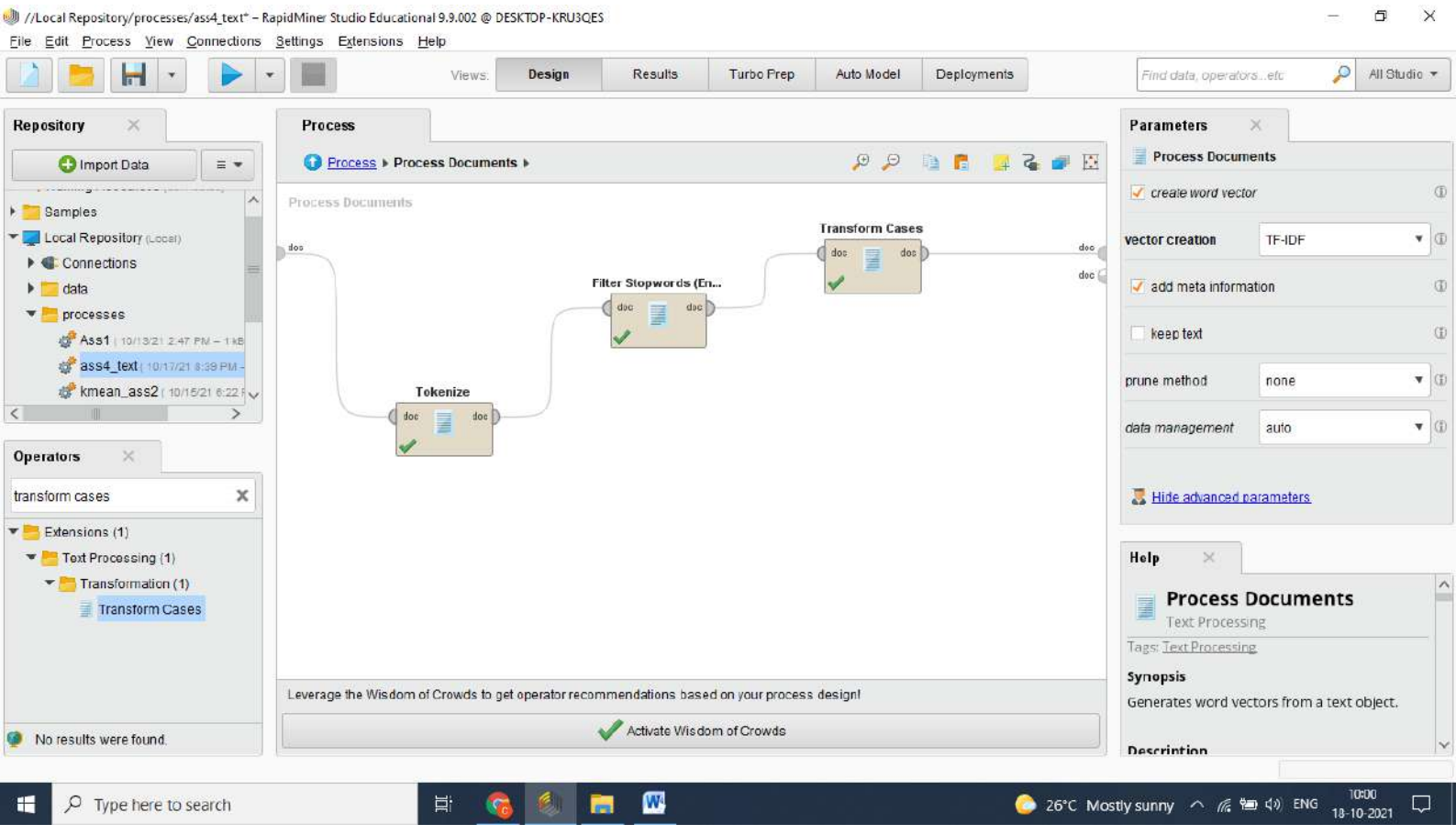
Double Click the Process Documents operator to get inside the operator. This is where we will links operators together to take the entire text document & split it down into it word components. This consists of several operators that can be chosen by going into the Operator panel & ~~so~~ looking at the Text Processing folder.

Now we are ready to filter the bag of words. In 'Filtering' folder under the 'Text Processing' operator folder, you can see the various filtering methods that you can apply to your process.

Conclusion:-

We are now able to see a word list containing all the different words in your document & their occurence count next to it in the 'Total Occurence' column. If you do not get this output, make sure that all of your nodes are connected correctly & also to the right type. Some errors are because your output at one node does not match the type expected at the i/p of the next node of an operator.


//Local Repository/processes/ass4_text* – RapidMiner Studio Educational 9.9.002 @ DESKTOP-KRU3QES
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Samples
Local Repository (Local)
Connections
data
processes
Ass1
ass4_text
kmean_ass2
Operators
transform cases
Extensions (1)
Text Processing (1)
Transformation (1)
Transform Cases
No results were found.
Process
Process
Read Document
Process Documents
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Process Documents
create word vector
vector creation
TF-IDF
add meta information
keep text
prune method
none
data management
auto
Hide advanced parameters
Help
Process Documents
Text Processing
Tags: Text Processing
Synopsis
Generates word vectors from a text object.
Description


//Local Repository/processes/ass4_text* – RapidMiner Studio Educational 9.9.002 @ DESKTOP-KRU3QES
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Samples
Local Repository (Local)
Connections
data
processes
Ass1
ass4_text
kmean_ass2
Operators
transform cases
Extensions (1)
Text Processing (1)
Transformation (1)
Transform Cases
No results were found.
Process
Process ▸ Process Documents ▸
Process Documents
Tokenize
Filter Stopwords (En...
Transform Cases
doc
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Process Documents
create word vector
vector creation
TF-IDF
add meta information
keep text
prune method
none
data management
auto
Hide advanced parameters
Help
Process Documents
Text Processing
Tags: Text Processing
Synopsis
Generates word vectors from a text object.
Description