Abdhel Muthazz Mashood
Student ID : 21595621

Business Intelligence Technologies: Practical 5: Text Mining

Weighting: 4% Due date: 23:59pm Friday 6th January 2023

A hotel has collected feedback from its customers and would like you to do some text mining to see what insights can be extracted from this data, eg are customers generally happy or not? If not, what kind of things are customers unhappy about? A manager has gone in and annotated each feedback with a 'positive' or 'negative' classification and has asked whether it would be possible to put together a model that could classify feedback in the future, based on this test data. The manager has also heard that 'sentiment analysis' can be undertaken on this data and wonders whether this might be useful.

Using the hotelfeedback dataset, carry out **two** data mining processes on this data. Annotate **each operator** to explain what it does, what parameters you are using and why, and then provide an analysis of the output.

(Processes may be classification, association rule mining, sentiment analysis, classification, clustering etc.)

Assessment Criteria:

Correct and documented processes following the CRISP methodology, showing sound understanding of the text mining methods used: 2 marks
Documentation and evaluation of results: 2 marks

Attribute information:

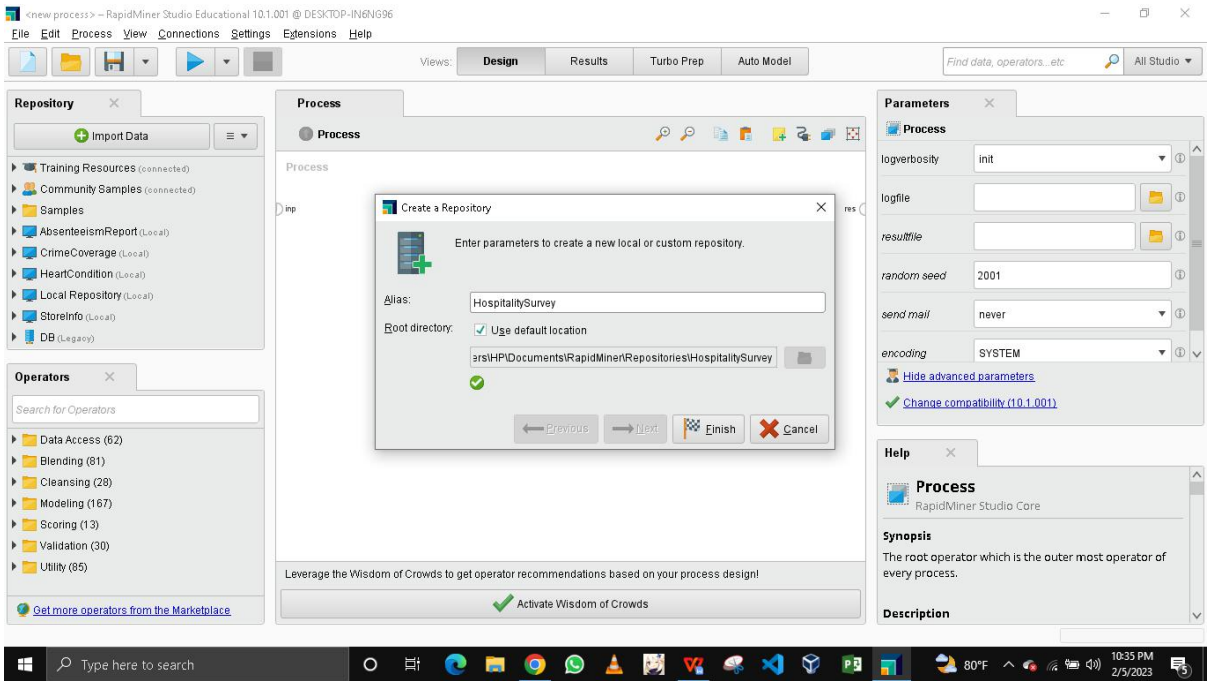
The data is contained in the 'hotelfeedback' excel file which contains two columns,

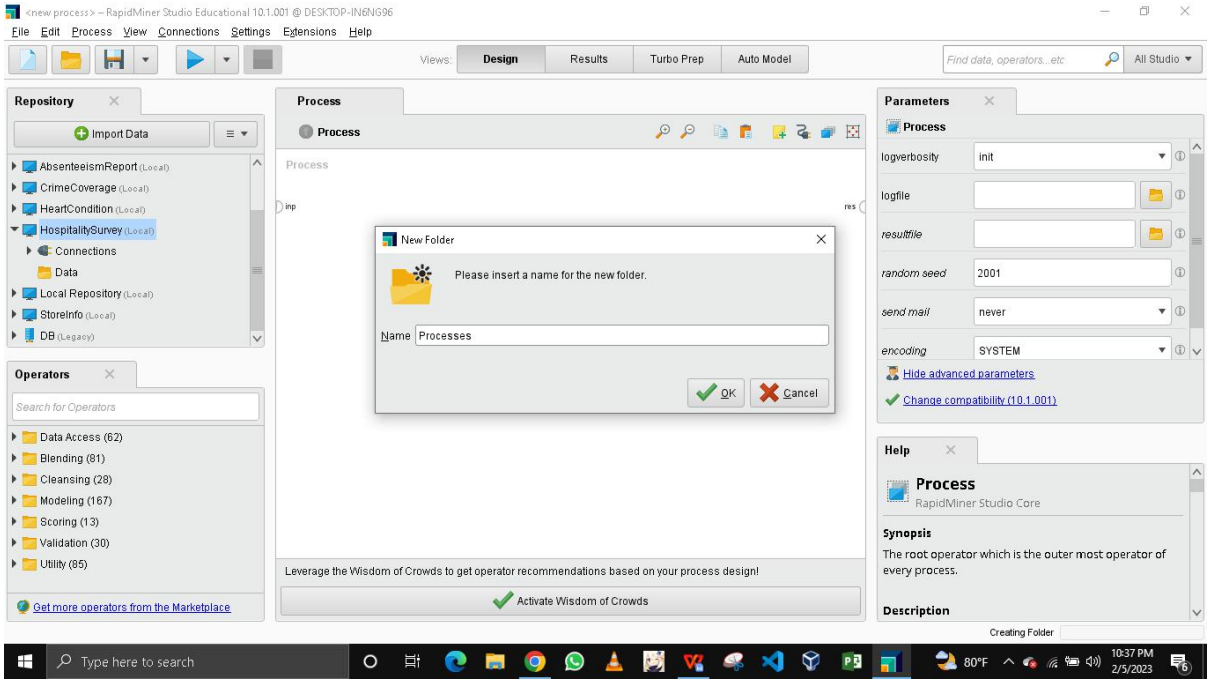
Column1: contains the classification of the feedback (either positive or negative)
Column2: contains the feedback text

Abdhel Muthazz Mashood
Student ID : 21595621

Practical Documentation

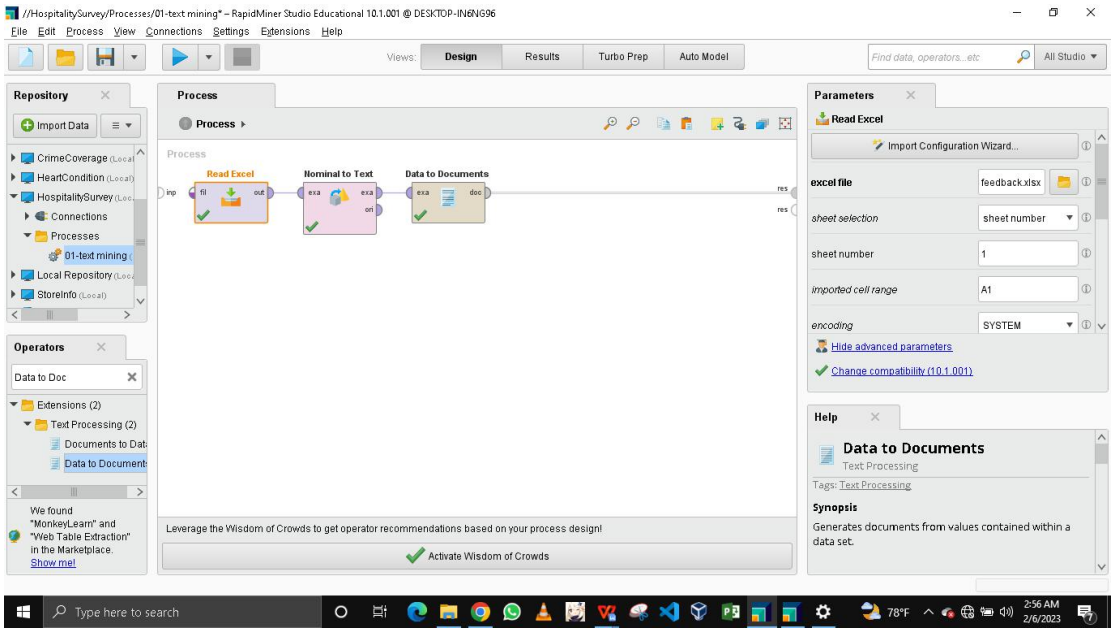
1. Created a Repository labeled “HospitalitySurvey” and created a subfolder named “Processes”

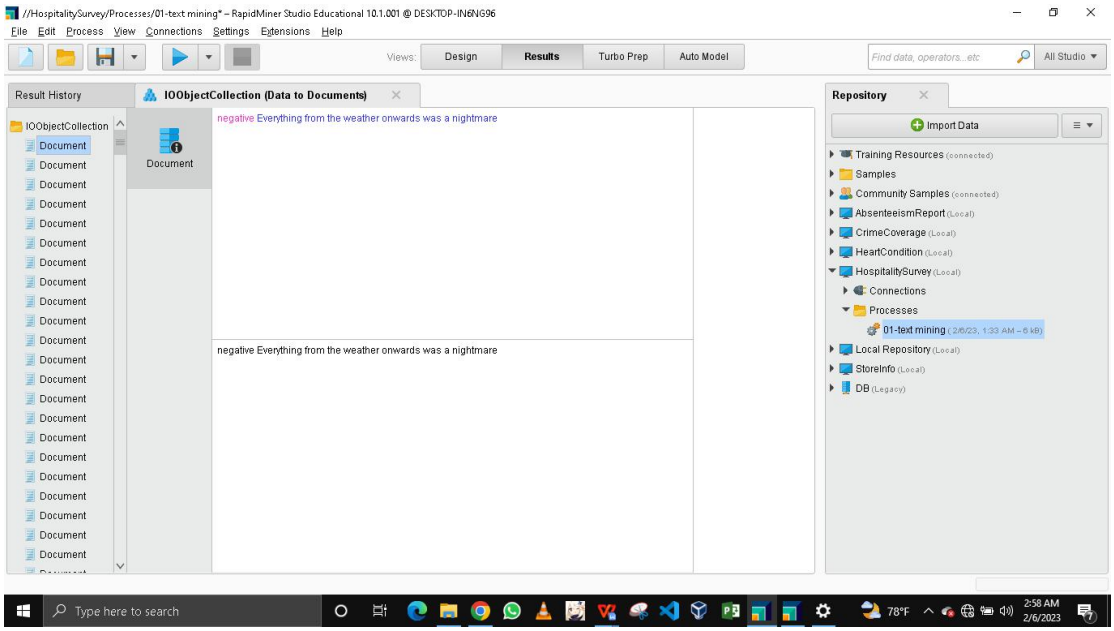




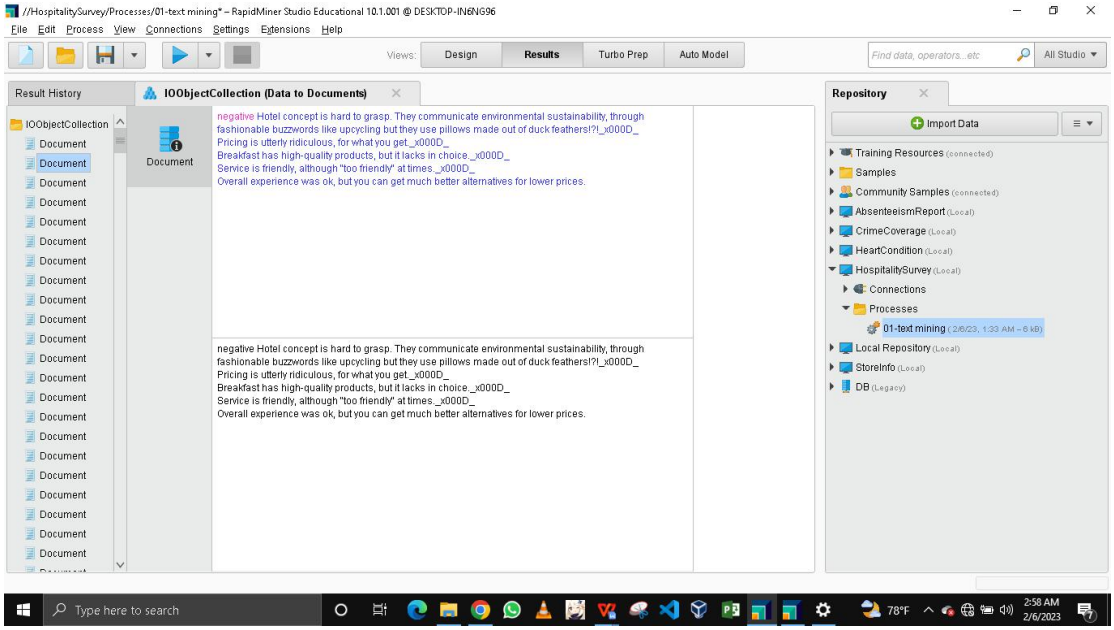
Abdhel Muthazz Mashood
Student ID : 21595621

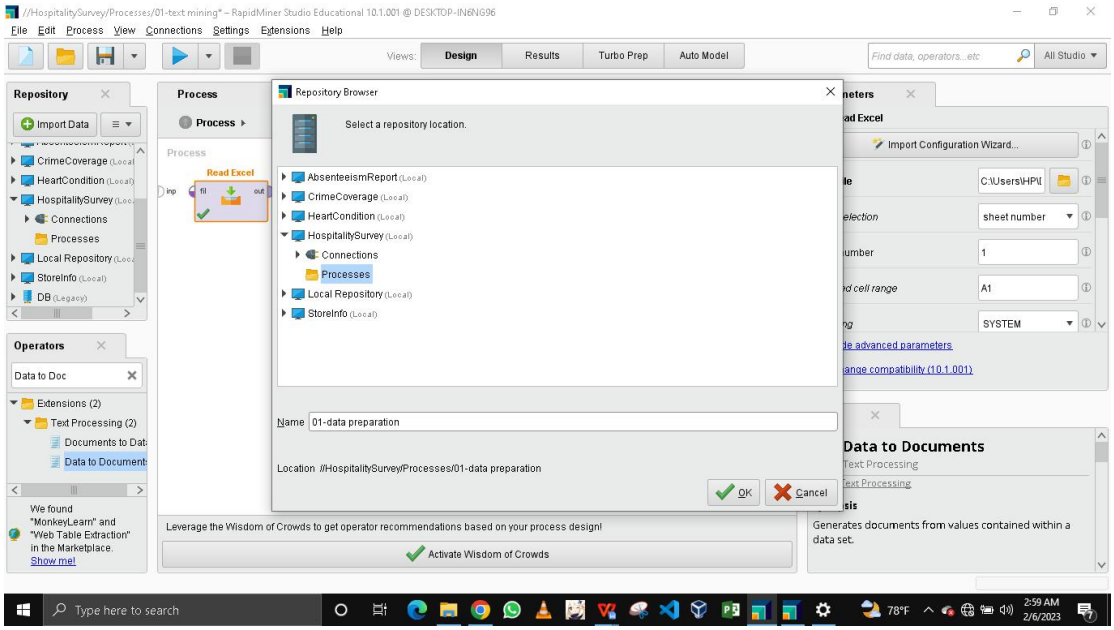
2. Used the “Read Excel” operator to read the hotelfeedback table and connected it to the “Nominal to Text” and “Data to Documents” operators respectively and saved the process as “01-data preparation”, given the customer feedback is stored in an excel sheet and needs to be prepared in the form of text documents prior to being processed for text mining





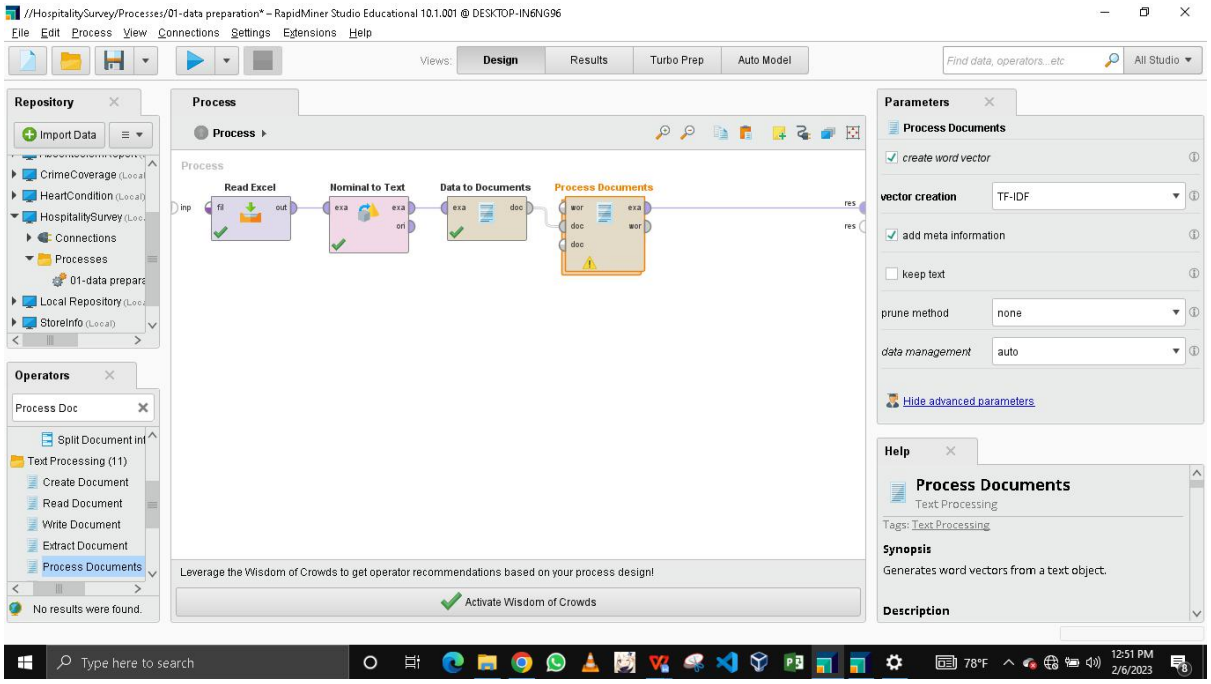
Abdhel Muthazz Mashood
Student ID : 21595621

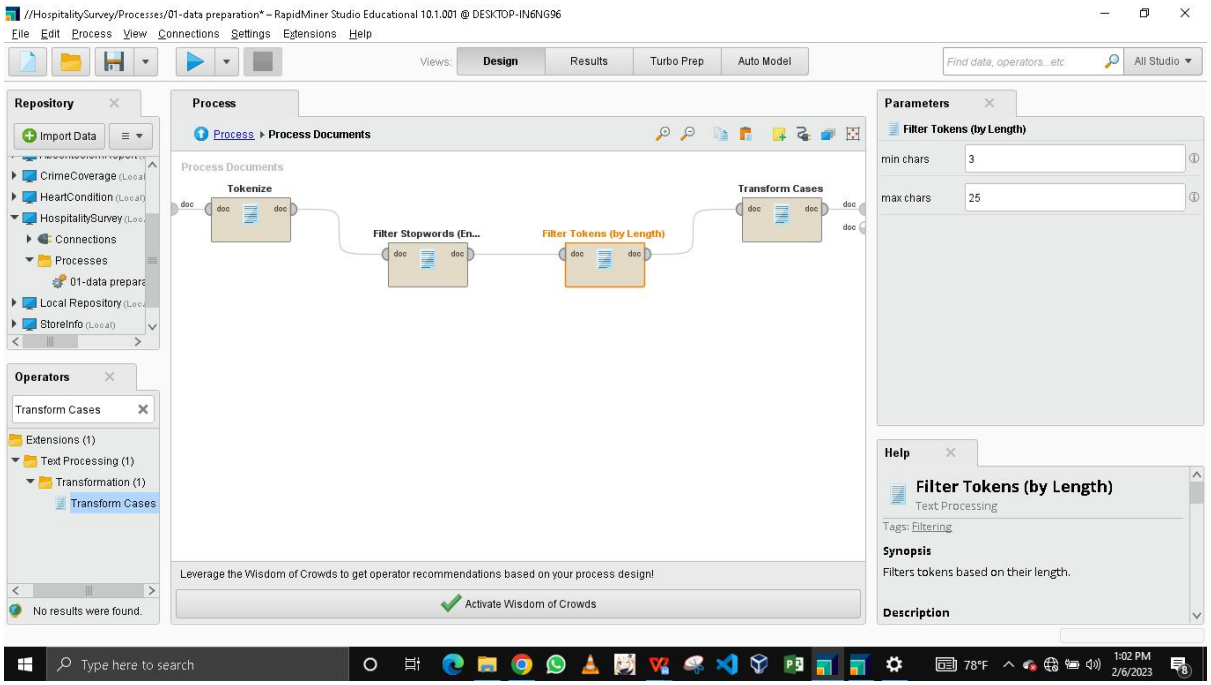




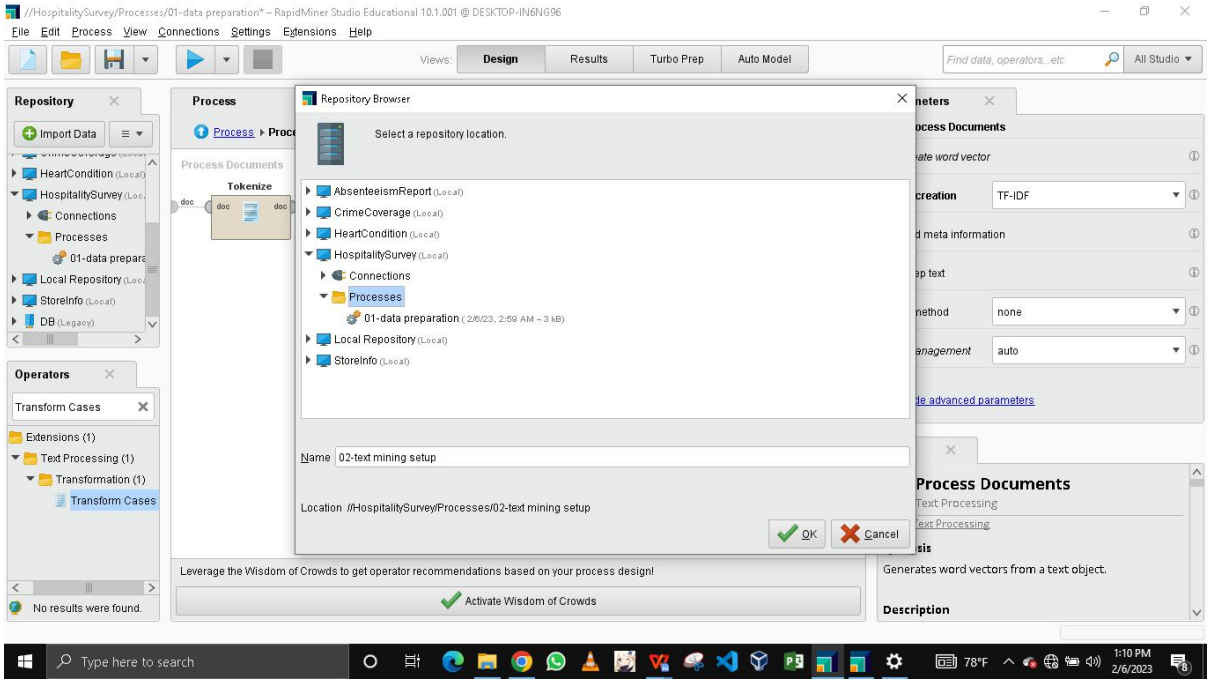
Abdhel Muthazz Mashood
Student ID : 21595621

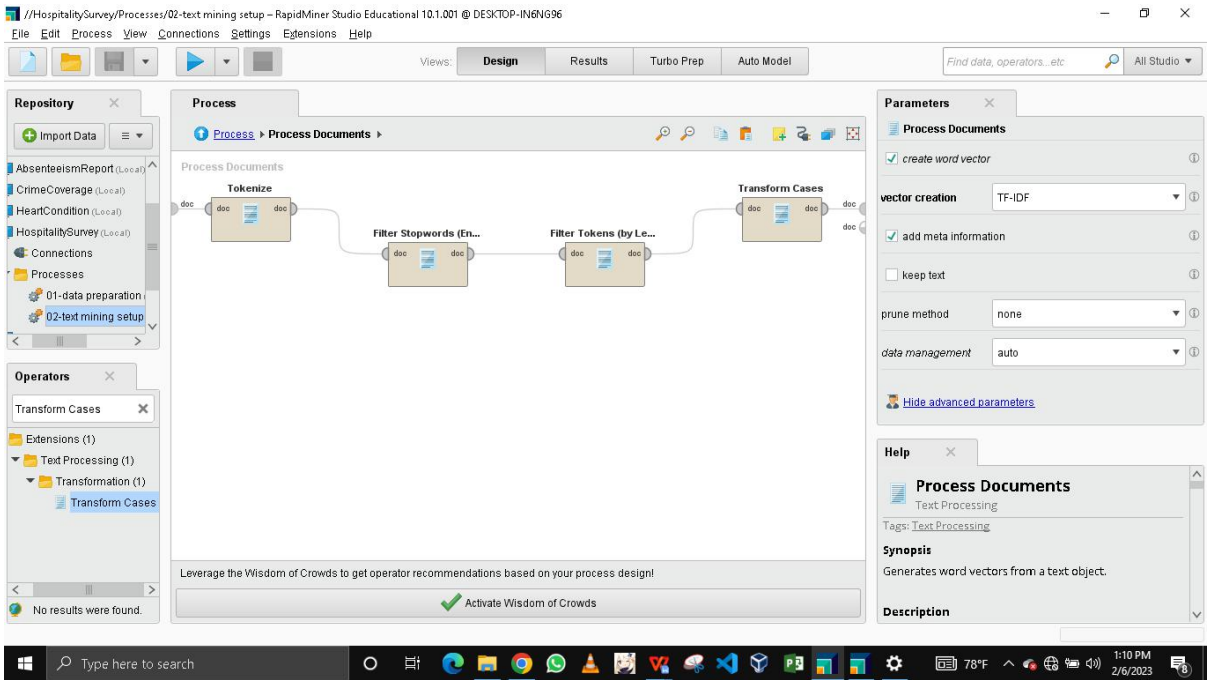
3. Fed the prepared data into the “Process Documents” operator which opens up into a sub-process window that should consist of the following text mining operators such as “Tokenize”, “Filter Stopwords (English)”, Filter Tokens (by Length)” and “Transform Cases” respectively and to set up the text mining process and minimize the recurrences of the same words in different cases; saving the process as “02-text mining setup”





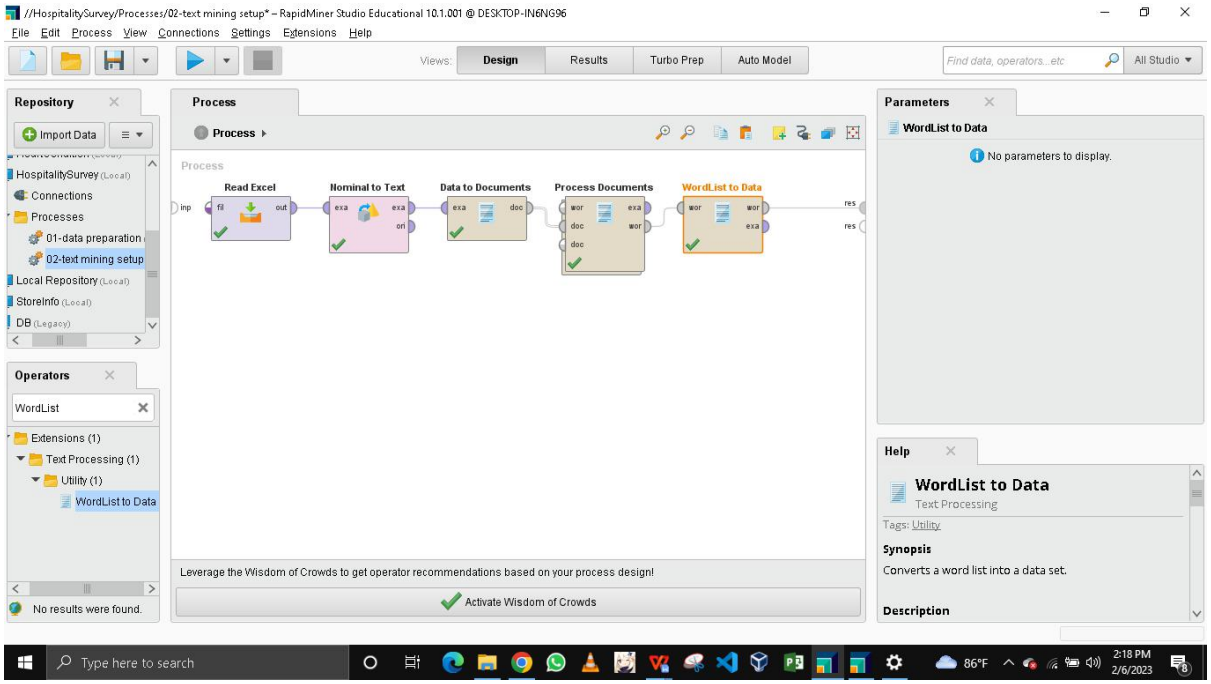
Abdhel Muthazz Mashood
Student ID : 21595621

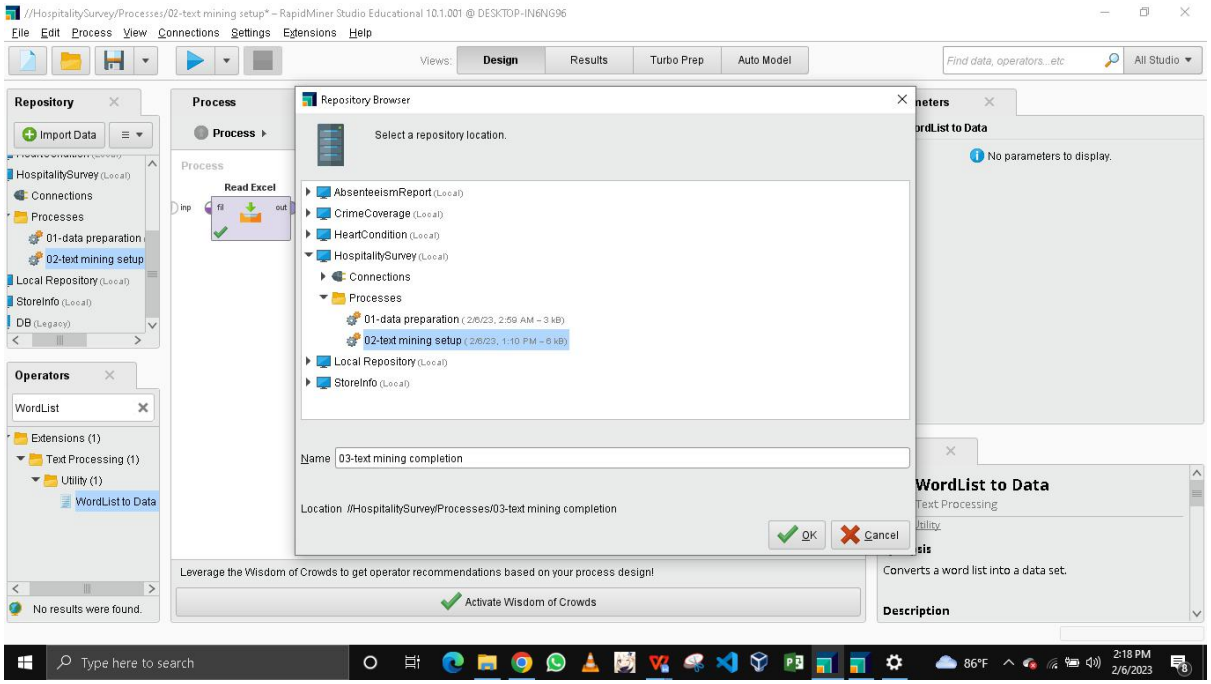




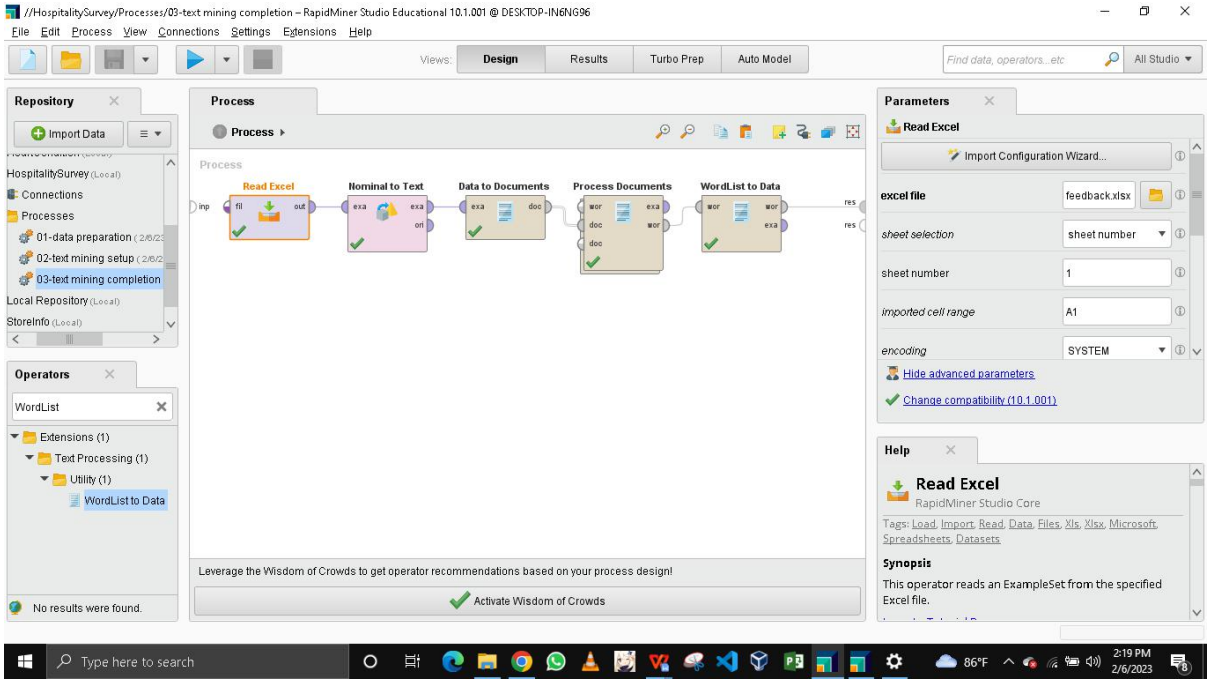
Abdhel Muthazz Mashood
Student ID : 21595621

4. Exited the sub-process window and passed the mined text through the “WordList to Data” operator to get a tabulated version of the text mining results.

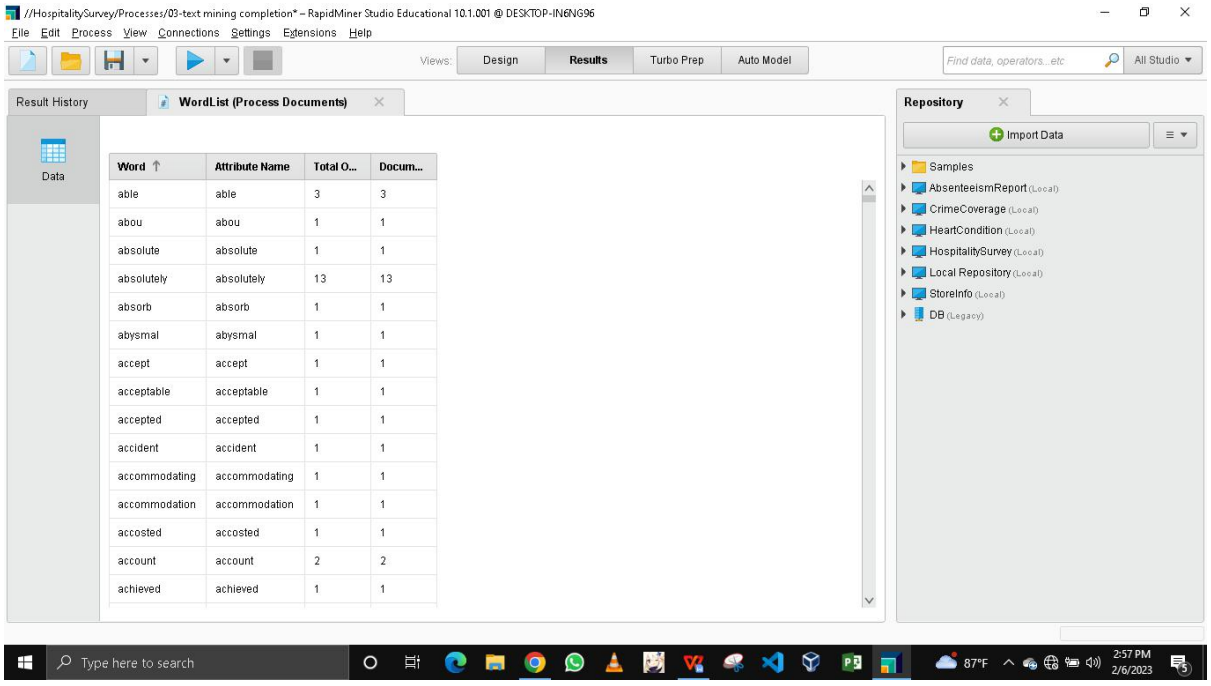




Abdhel Muthazz Mashood
Student ID : 21595621

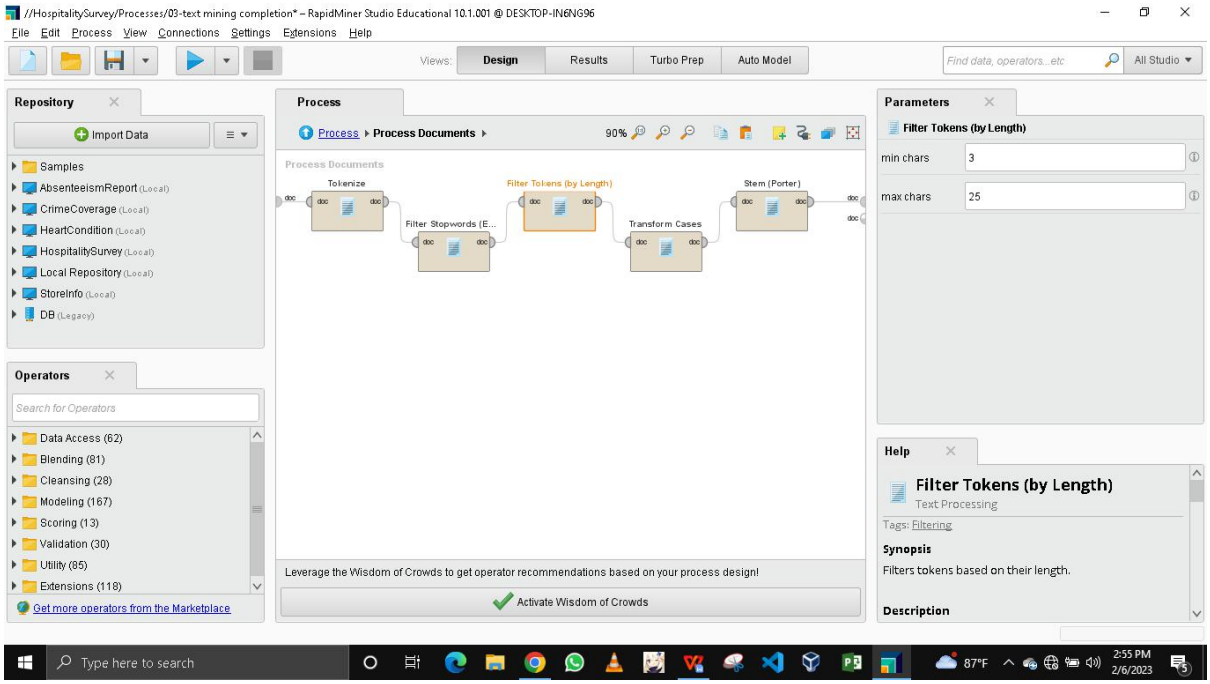


Text Mining Results

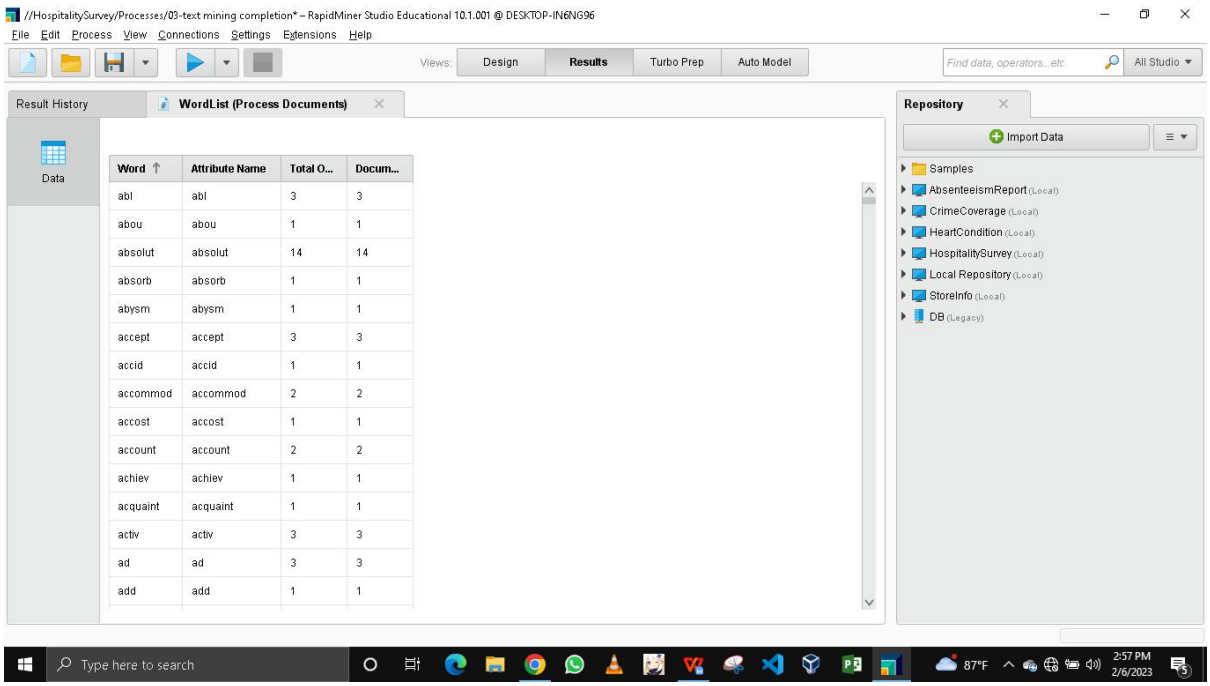


Abdhel Muthazz Mashood
Student ID : 21595621

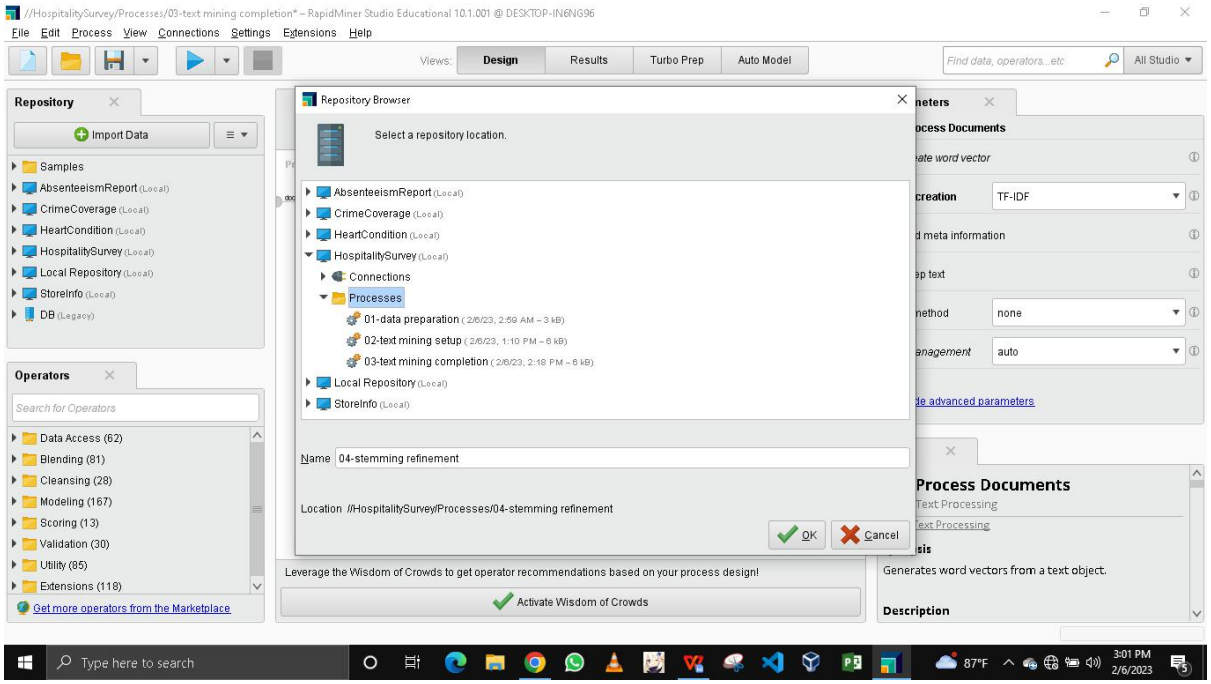
5. Refined the process through the addition of the “Stem (Porter)” operator within the sub-process window of the “Process Documents” operator and saved the process as “04-stemming refinement”

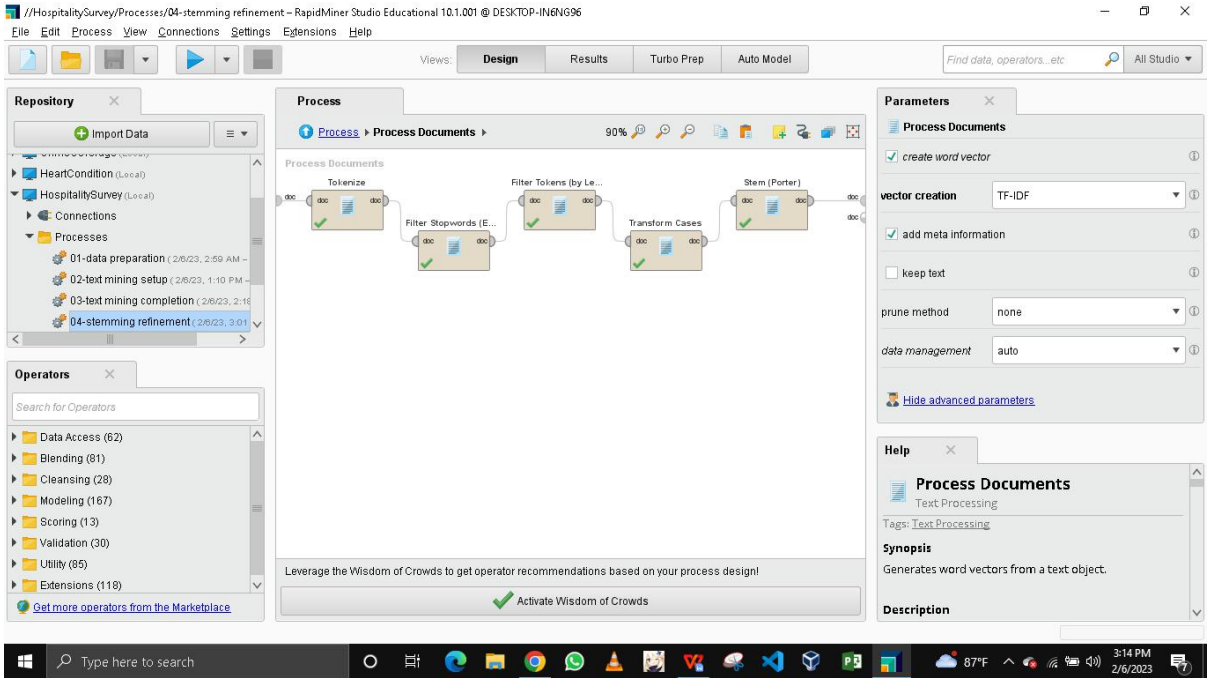


Compared to the text mining results above the results of the refined process using the “Stem (Porter)” operator contains reduced recurrences of words with common root words as a single term such as the total occurrences of the words “absolute(1)” and “absolutely(13)” as a shared total occurrence of the root word “absolute(14)”



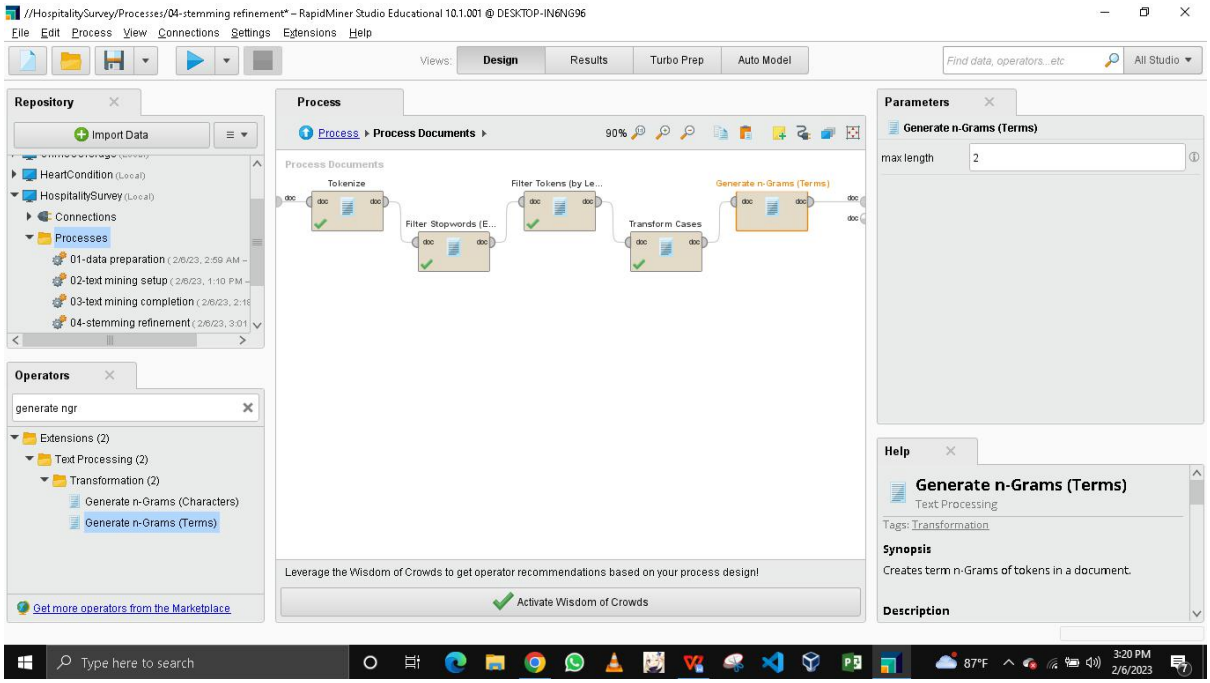
Abdhel Muthazz Mashood
Student ID : 21595621



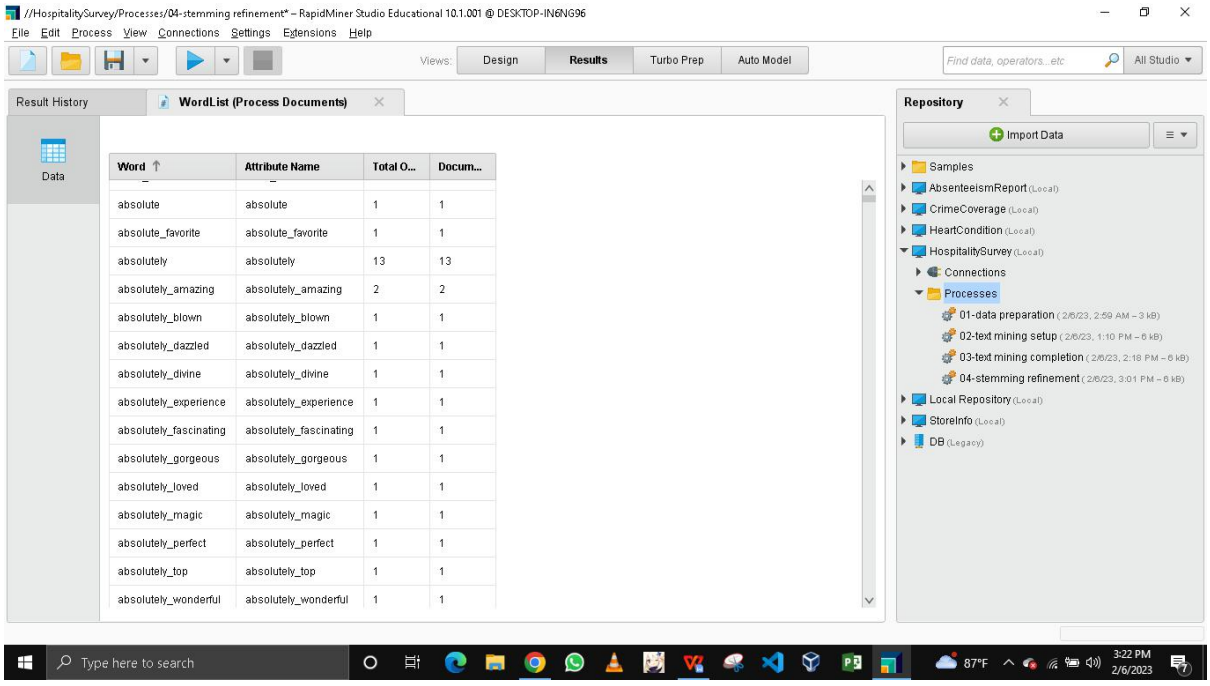


Abdhel Muthazz Mashood
Student ID : 21595621

5. Edited the previous process by replacing the “Stem (Porter)” operator with the “Generate n-Grams” operator within the sub-process window of the “Process Documents” operator and saved the process as “05-n-grams refinement” respectively



Given the max length limitation of two word generation in the “Generate n-Grams (Terms)” operator, shown below are the results of the context in which the word “absolute” is attached to in order minimize the misinterpretation of the feedback which in this case can be concluded that the word “absolute” appears in generally positive reviews of the hotel



Abdhel Muthazz Mashood
Student ID : 21595621