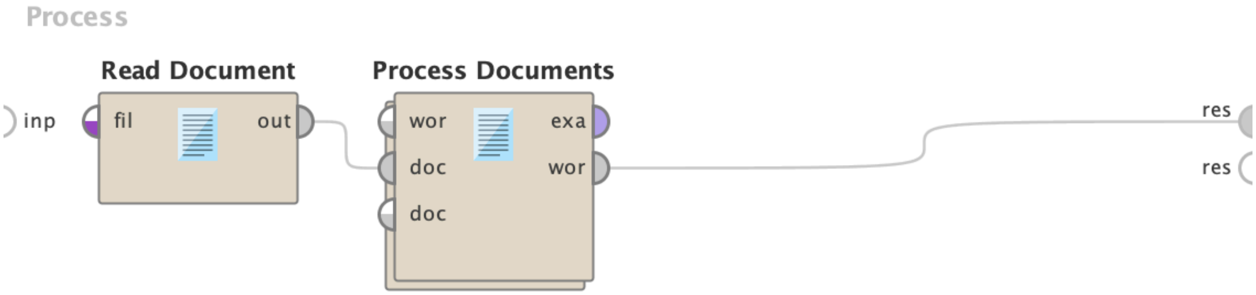
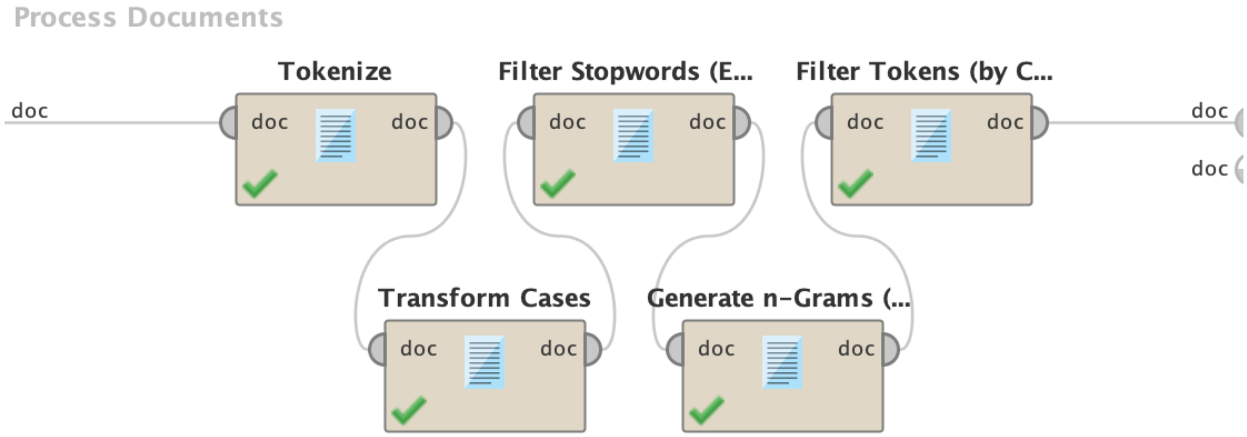
Assignment 5 Section 1

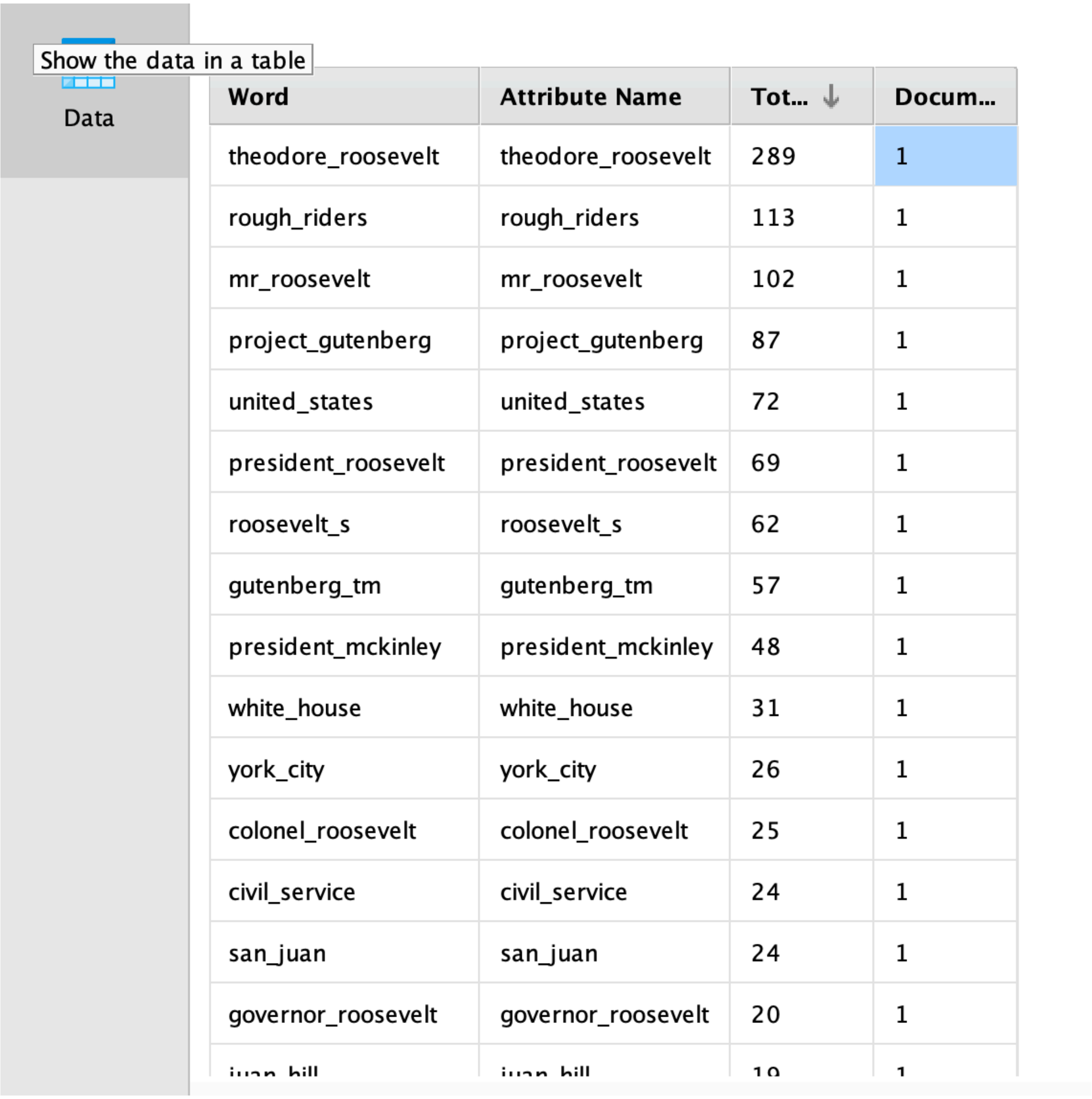
1. Get Read Document and Process Document operators.
   The file used here is *American Boy's Life of Theodore Roosevelt* by Edward Stratemeyer



2. In the Process Document operator, utilize the following operators to analyze the document.
   Tokenize: Breaks up document into words, removes punctuations.
   Transform Cases: converts all letters into lowercase to avoid redundancy in counting.
   Filter Stopwords: filters propositions, articles and words with no meaning
   Generate n-Grams (max length-2) connects max of two words with an under score.
   Helps determine the probability that a word will appear after another.
   Filter Token (by content): filter words with a particular character; in this case '_'.



3. Our result shows that there were 289 instances where the words, theodore and roosevelt appear in this order. Furthermore, prases like, ‘mr roosevelt’ or ‘president roosevelt’ are among the words that appear frequently. Phrases such as ‘rough riders’ and ‘united states’ also appear significantly many times. The use of the phrase ‘rough riders’ would be consistent with the organization name that Theodore Roosevelt and his friend Dr. Leonard Wood founded.



Word	Attribute Name	Tot...	Docum...
theodore_roosevelt	theodore_roosevelt	289	1
rough_riders	rough_riders	113	1
mr_roosevelt	mr_roosevelt	102	1
project_gutenberg	project_gutenberg	87	1
united_states	united_states	72	1
president_roosevelt	president_roosevelt	69	1
roosevelt_s	roosevelt_s	62	1
gutenberg_tm	gutenberg_tm	57	1
president_mckinley	president_mckinley	48	1
white_house	white_house	31	1
york_city	york_city	26	1
colonel_roosevelt	colonel_roosevelt	25	1
civil_service	civil_service	24	1
san_juan	san_juan	24	1
governor_roosevelt	governor_roosevelt	20	1
juan_hill	juan_hill	19	1